United States
Department of
Agriculture

Service Center
Implementation
(SCI)

Implementation of Geospatial Data Warehouses

Prepared by
Science Applications International Corporation (SAIC)
for
The Data Management Team #5 on Geospatial Data Standards

Abstract: This document provides the results of an evaluation of alternatives and recommendations for implementing a geospatial data warehouse framework for the USDA Service Center Agencies. The plan includes a discussion of the technical architecture, physical location of warehouses, operational requirements, resource and staff needs and budgets. The objective of this document is to present an implementation path that begins with the current architecture and provides growth path that meets the near-term requirements and to be sustained through the long-term vision of geospatial data distribution in USDA farm agencies.

Keywords: geospatial, data management, architecture, implementation

Introduction

(This introduction is not part of the Geospatial Data Requirements Document)

The Service Center Initiative (SCI) Data Management Team #5: Geospatial Data Standards developed the September 2000 *Implementation of Geospatial Data Warehouses* to illustrate the USDA vision of how geospatial data warehousing will be used for on-line data access and data dissemination in the near and long-term.

The individuals who contributed to the development of this guide are:

David Anderson, (NRCS) Service Center Data Team Leader
Carol Ernst, (FSA) Co-Leader
Emil Horvath, (NRCS) Co-Leader

Tommy Parham (NRCS, Dir. NCGC)
Ron Nicholls (FSA, Dir. APFO)
Dennis Lytle (NRCS)
Kent Williams (FSA)
Steve Nechero (NRCS)
Kenny Legleither (NRCS)
Larry L. Davis (NRCS)
Dwain Daniels (NRCS)
Elaine Ortiz (NRCS)
Elizabeth Young(NRCS)
Nicole Soltyka (SAIC)
Tom McCarty (SAIC)
Randy Frosh (Unisys)

Figure 0-1 – Working Group List

TABLE OF CONTENTS

List of Figures

List of Tables

Executive Summary and Recommendations

In 1998 the Service Center Agencies (FSA, NRCS and RD) established a vision and goals for a unified geospatial data management framework and geospatial data warehouse. The Service Center Modernization Initiative efforts of the United States Department of Agriculture (USDA) Service Center Agenc ies (SCA) has identified that the ability to manage digital spatial information has a high payoff for improving SCA ability to meet business needs. Management of spatial information is accomplished with Geographic Information System (GIS) technology. The implementation of GIS depends upon the ability to obtain, manage, and deliver geospatial data in an efficient and cost-effective way. The *USDA Service Center Geographic Information System (GIS) Strategy, August 1998* identified 23 separate geospatial data sets that are needed by Service Center Staff. Subsequent requirement analysis has identified even more geospatial themes used at the local, state, and national levels.

Service Center Agencies Lack the Capacity to Meet Anticipated Demand for Geospatial Data and Information – With existing processes the acquisition, integration, distribution to 2600 offices and the public, update and maintenance are difficult if not impossible. Agencies do not have adequate infrastructure or resources to manage and deliver geospatial data to Service Centers, partners and customers. NRCS state offices are attempting to also deliver geospatial data to Service Centers. Some have developed agreements with universities or state agencies. In some cases these processes duplicate processes at NCGC. In other state offices NRCS does not have resources for geospatial data distribution forcing the state to delay GIS implementation or to rely on Service Center staff who are usually less knowledgeable of the process and requirements. In both of these cases costs are increased and productivity is decreased.

The NRCS Customer Service Toolkit is being implemented in more than 45 states. Toolkit requires at a minimum orthoimagery and digital soils data. More than 2000 field conservationists will be equipped with ArcView desktop GIS in early FY 2001, but do not have ready access to the geospatial data layers that will help them do their job. Data delivery is a fundamental program requirement. Five years after the Chief's blue ribbon task force report, the NRCS is still data rich, but information poor.

FSA has the responsibility for developing, managing, and delivering enhanced orthoimagery. This imagery is the primary base for digitizing other geospatial data themes. Customers have also expressed an interest in obtaining this imagery for their own GIS systems. FSA is very committed to digitizing Common Land Units (CLU). CLUs are the basis for the administration for many of their programs and considered one of the basic layers for Service Centers. By the end of FY-2000 FSA had CLU's and the ability to maintain them in 200 counties. They currently have digitizing centers in 7 states and in FY-2001 will be expanding the sites in 3 of these states.

Both FSA and NRCS lack the infrastructural capacity to provide current data to internal users and ready access to data by the public.

Data Warehouses, Web Applications, and Automated Data Delivery - With the growing maturity of web technology and improved bandwidth, the time is right for the agencies to deliver information on a new technological foundation. This new foundation consists of on-line data warehouses as the authoritative source of data, web applications that utilize the data via the Internet and automation of the data dissemination process. This approach will minimize the manual effort required to disseminate and manage the data. It also will meet the objective presented in the *Geospatial Data Acquisition, Integration and Delivery National Implementation Strategy Plan, September 1999* to deliver geospatial data to Service Center offices as a turnkey process that minimizes the data management task at the local level.

Scope and Objective

The vision is coordinated data warehouses that provide on-line and seamless access and delivery of geo data for internal business processes, agency business partners, and external customers.

Implementation of Geospatial Data Warehouses, October 2000 describes in detail a plan for implementing these warehouses. The plan provides alternative models for implementation. It describes the technical architecture (servers, telecommunications, data management software, etc.); physical location of the warehouses, operational requirements, metadata management requirements, resource and staff needs and budgets.

This document was prepared with support from Science Applications International Corporation (SAIC) for the Service Center Modernization Initiative, Data Management, Geospatial Data Standards Team. The Geospatial Data Standards Team is made up of individuals who have responsibility for managing geospatial data within NRCS and FSA. The use of GIS for Rural Development business applications is in its infancy. As this need matures RD will become a more active player in managing geospatial data and may develop its own resources for managing data or use the existing resources of FSA or NRCS.

The vision is a set of coordinated data warehouses that provide on-line and seamless geo data access and delivery for delivery to Service Centers and external customers.

Data Management Alternatives Evaluated – Two major documents were created by this effort. Phase I, *Geospatial Data Requirements, April 2000* identified and described the geospatial data sets to be managed, the physical characteristics for each data set, and the requirements for managing the data. *Implementation of Geospatial Data Warehouses, October 2000* represents the second phase and provides a detailed implementation plan for the best approach for managing geospatial data. It develops a comprehensive comparison between various alternatives for managing agency geospatial data assets. The alternatives included storing all the geospatial data in a single central data repository, and various options for distributing the data management function at multiple locations. Each alternative was evaluated against estimated costs of implementation, operations (hardware, software, telecommunications, training, and staffing); and performance, as well as intangibles such as ease of implementation, likelihood of success, and sustainability.

After initial comparison of all the alternatives the team focused on options of a single data warehouse, and distributing data at the current Data Acquisition and Integration Centers (DAI) - FSA Aerial Photography Field Office, (APFO) at Salt Lake City, UT and NRCS National Cartography and Geospatial Center, (NCGC) at Fort Worth TX. The estimates that were developed indicate that the cost for implementing a central facility is about $2,136,000, while the cost for distributed data management is about $2,702,000, when the existing infrastructure (servers, software, staff,) in the current DAI Centers was factored in the actual costs. This represents an actual cost difference of about $566,000. These cost figures represent an initial deployment level and should cover the costs required through FY 02.

Recommendations

Although there would be some cost advantage for centralizing data management at a single location, the cost difference is not sufficient to justify the development of a single large center for data distribution. Therefore, the data management team recommends that a distributed model for managing data be employed. Dissemination of geospatial data would be accomplished through a network of distributed warehouses including the two existing USDA Service Center Agency DAIs and potentially other USDA, non-USDA, local, state, regional and national nodes that host geospatial data sets. These warehouses should be linked by a common Internet portal that provides a one-stop-shopping and data ordering services making the distributed nature of the data appear seamless to the users.

The operation and maintenance of the distributed warehouses will rely heavily on partnerships and leveraging the resources of the DAIs, Electronic Access web farms, and agency Information Technology organizations.

1) Establish on-line Data Warehouses at the Data Acquisition and Integration Centers

The primary DAIs, APFO and NCGC, should establish and maintain the data warehouses that are the authoritative source for data dissemination and on-line applications. The specific data housed by each organization has been established by prior agreement. These centers are responsible for acquisition, integration, storage, archival, maintenance, and dissemination of geospatial data to internal users and the public. The security infrastructure required to meet Department security policy will be implemented at the DAIs to provide for public access to data. The data ordering and dissemination function will be automated to minimize the manual effort required. As other data sources are brought on-line other data centers may be established in the future.

NRCS should focus on NCGC as its primary data center and build capacity to electronically deliver the geospatial data for which it is responsible to its internal and external customers. It should also build capacity at the Fort Collins web farm to support web applications that use geospatial data. The agency should build a single integrated natural resource data warehouse as the authoritative source for geospatial data at the Fort Worth data center.

FSA should build capacity at APFO to manage and provide on-line access to imagery data, and the Kansas City web farm to support their web business applications. APFO and US Forest Service Geometronics Service Center are co-located and are working to develop a cost share arrangement to increase bandwidth. FSA has not established how it will manage other business data sets such as the Common Land Unit and Land Use data.

2) On-line Web Applications are Housed at Electronic Access Web Farms

The USDA Service Center Agencies are building web farms in Kansas City MO., Fort Collins CO, and Saint Louis MO, to provide USDA web products and services internally and to the public. These web farms have implemented the infrastructure necessary for maintaining the agency's presence on the Internet and for meeting the emerging requirements of the new "e-government" laws. Each of the web farms provide: high speed internet access to both the Internet and the USDA intranet; robust security features including firewalls, intrusion detection, vulnerability monitors, and user authentication; common web services including index/search, web conferencing, and discussion group administration; dedicated resources including staffing and funds to ensure appropriate operations; and consistent policies and procedures to ensure dependable delivery of USDA's products and services.

The infrastructure being put in place for these web farms will be leveraged to the fullest extent possible. Web farms should host on-line public and Intranet web applications. Request for geospatial data would be processed at a web farm through a common data access portal that provides one-stop-shopping and data ordering services. Requests for data are routed to the appropriate DAI for processing the orders and delivery. The requests, order placement, packaging, and delivery processes will be automated to be performed with little or no manual intervention. On-line applications that access and process information would be hosted within the web farm. Access to geospatial data at the DAI will be through high-speed telecommunication links or, where this is not feasible, the data set would be mirrored at the web farm.

Service Center Agencies have agreed that Fort Collins Web Farm will provide a common point of on-line entry into the data repositories residing at Fort Worth (NCGC) and Salt Lake City (APFO). As envisioned the Lighthouse/Gateway project will eventually provide on-line access to data for applications, such as RUSLE II and watershed runoff models. NCGC and APFO will continue to be a main source of data for Service Center Agency GIS users, through the FTP sites and/or data on CDs and tape, in the foreseeable future. On-line delivery of data is not only desirable but necessary, yet is currently hampered by insufficient bandwidth between data repositories and service centers.

Agency Information Technology organizations would be responsible for developing, supporting, and maintaining web applications and the common data access portal which access the distributed data warehouses. IT organizations will also support the IT functions within the DAI and manage the telecommunications network.

3) Establish a Common Internet Portal as a "One-Stop-Shopping" Service for Geospatial Data

The distributed nature of the data will appear seamless to the users by linking the data warehouses through a common Internet portal that provides one-stop-shopping and data ordering services. Those who need geospatial data currently must visit many web sites, each having their own look-and-feel, navigation, and format, or make many phone calls to locate the information they need. In some instances once the data are located they must wait days for it to be delivered. The data is distributed across many servers, many data centers, and managed by many different organizations.

The common portal will provide easy access a to geospatial data and information to internal and external customers including the general public. Providing a common access portal is critical to timely and efficient delivery of data. The common portal would allow the user to identify a location on an on-screen map then see what data exists for that specific location. The users are then sent to the existing web sites for the data they select or place an order for the data on-line. The single access point would not require that the existing web sites or processes for delivering data be replaced, but would provide a common access point for locating that data.

A primary portal for USDA geospatial data will be directly accessible from the agency's home page. It will then link seamlessly to various programs, data, applications, and on-line databases. Examples are WCC's climate and ITC/NSSC's NASIS to NCGC's FTP sites, and APFO orthophotography. A primary portal must be flexible enough to accommodate users who know where data they need resides, allowing them seamless, direct access to DAI sites. This would be the primary method of access to the data, but alternate methods could be developed as applications or users needs change.

Specific Needs

Below are the specific needs for establishing data warehouses and on-line access to data.

APFO – Increase server capacity and on-line storage (8TB) $823,458
2 T1 Lines to USDA Backbone at Kansas City $37,500
Increase staff by 1.5 positions $231,733 (contract costs)
Implement Department security measures $200,000

NCGC – Increase server capacity, on-line storage (5 TB) $577,546
Additional software $34,000
2 T1 Lines to USDA Backbone at Kansas City $26,600

Increase staff .3 positions $39,811 (contract costs)
Implement Department security measures $32,000

Web Farm (Fort Collins)

Increased server capacity to house the common data access portal and on-line applications
$366,750
Additional Software $66,000
Increased staff 1.75 positions $266,282 (contract costs)

1. Introduction

Since its inception, the USDA has relied on geospatial information to accomplish its mission. In the past 30 years, the availability of digital geospatial data and tools created efficiencies in the quality of service that can be provided to USDA customers. Many successes have resulted from the individual efforts of the Service Center Agencies (FSA, NRCS and RD). In 1998, these agencies established a vision and goals for a unified geospatial data management framework and geospatial data warehouse (GDW). This plan is one of a series of steps taken to describe and implement the USDA Service Center GDW.

2. Acronyms

AID	Acquisition, Integration and Delivery
APFO	Aerial Photography Field Office
AS	Application Server
BPR	Business Process Re-engineering
CCE	Common Computing Environment
CD-ROM	Compact Disk Read Only Memory
CLU	Common Land Unit
CST	Customer Service Toolkit
CRP	Conservation Reserve Program
CARAA	Conservation Area Resource Analysis and Assessment
DAI	Data Acquisition and Integration
DBMS	Data Base Management System
DEM	Digital Elevation Model
DHTML	Dynamic Hyper Text Markup Language
DLG	Digital Line Graph
DOQ	Digital Orthophotography Quadrangle
DOQQ	Digital Orthophotography Quarter Quadrangle
DRG	Digital Raster Graph
DS	Data Server
EA	Electronic Access
EROS	Earth Resource Observation System
EDC	EROS Data Center
EAI	Electronic Access Initiative
ETL	Extraction, Transformation, Load
FAS	Foreign Agricultural Service
FSA	Farm Service Agency
FTE	Full Time Equivalent
FTP	File Transfer Protocol
FY	Fiscal Year
GDW	Geospatial Data Warehouse
GIS	Geographic Information Systems
GPS	Global Positioning Systems
HTML	HyperText Markup Language

HTTP	HyperText Transfer Protocol
ID	Identification
IIS	Internet Information Server
IP	Internet Protocol
IT	Information Technology
LAN	Local Area Network
MDOQ	Mosaicked Digital Orthophotography Quadrangle
MIME	Multi-purpose Internet Mail Extensions
NAPP	National Aerial Photography Program
NCGC	National Cartography and Geospatial Center
NCSS	National Cooperative Soil Survey
NRCS	Natural Resources Conservation Service
NRI	Natural Resources Inventory
O&M	Operations and Maintenance
RD	Rural Development
RDG	Resource Data Gateway
RFI	Request For Information
RFP	Request For Proposal
SC	Service Center
SCA	Service Center Agencies
SCI	Service Center Initiative
SCIMS	Service Center Information Management System
SCM	Service Center Modernization
SDE	Spatial Data Engine
SRS	Spatial Reference System
USDA	United States Department of Agriculture
USFS	United States Forest Service
USGS	United States Geological Survey
WAN	Wide Area Network
WDC	Washington, D.C.
WS	Web Server
WWW	World Wide Web

3. Background

The Business Process Reengineering (BPR) efforts of the United States Department of Agriculture (USDA) Service Center Agencies (SCA) has identified that the ability to manage digital spatial information has a high payoff for improving SCA ability to meet business needs. Management of spatial information is accomplished with Geographic Information System (GIS) technology. Implementation of GIS depends upon the ability to obtain, manage, and deliver geospatial data in an efficient and cost-effective way. The *USDA Service Center Geographic Information System (GIS) Strategy, August 1998* [A1], identified 23 separate geospatial data sets that are needed by Service Center Staff, four of which are identified as especially important and high priority. A Geospatial Acquisition, Integration, and Delivery (AID) Team further elaborated on the needs for geospatial data and as a result recommended that a Geospatial Data Implementation Plan be developed to guide the implementation of an infrastructure for physically managing the data. The AID Team presented their recommendations in the *Geospatial Data Acquisition, Integration and Delivery National Implementation Strategy Plan, September 1999* [A2]

4. Scope and Objective

The objective of this plan is to develop a Geospatial Data Implementation Plan for the USDA SCA. The geospatial Data Implementation Plan will provide a detailed roadmap for implementing coordinated processes and procedures and an infrastructure for managing geospatial data sets. The current vision is a set of coordinated data warehouses that provide on-line and seamless geo data access and delivery to SCA and external customers. This document will describe in detail a plan for implementing these warehouses. It will describe the technical architecture (servers, telecommunications, data management software, etc.); physical location of the warehouses, operational requirements, metadata management requirements, resource and staff needs and budgets.

The vision is a set of coordinated data warehouses that provide on-line and seamless geo data access and delivery for delivery to Service Centers and external customers.

The plan consists of two phases. The first phase, *Geospatial Data Requirements, April 2000* [A3] identified and described the geospatial data sets to be managed, identified the physical characteristics for each data set, and identified the requirements for managing the data. This second phase provides a detailed implementation plan that identifies the technical architecture and operational requirements.

5. Roadmap to the Vision

Advances in spatial information technology are converging to enable the Service Center Geospatial Data Warehouse (GDW) vision to come to fruition. The implementation plan must be managed to provide a flexible but focused implementation roadmap to meet the GDW vision. To understand the future direction, one must first review where the USDA SCI (Service Center Initiative) has been and the guiding forces that shaped where were are today. In 1998, the Service Center GIS Strategy [A1] was drafted to identify the primary goals and benefits of GIS, the critical and non-critical data themes needed for agency business, and budget guidance on how to achieve those goals. In the following fiscal year, a series of geospatial standards were drafted to help guide the acquisition of geospatial data for the Service Centers including:

- *Standard for Geospatial Data, January 2000* [A4]
- *Standard for Geospatial Data Set Metadata, August 1999* [A5]
- *Standard for Geospatial Dataset File Naming, August 2000* [A6]
- *Standard for Service Center Tabular Metadata, September 1999* [A7]

As recommended by the GIS Strategy [A1], several BPR pilot activities were initiated to re-engineer specific business processes. One primary goal of the pilots was to demonstrate the benefits of GIS to USDA Service Centers. The following table lists detailed information for those BPR pilots that used geospatial data and technology.

Table 5-1 - Service Center Geospatial BPR Pilots

Pilot Name	**Pilot Descriptions**
Customer Service Toolkit (CST)	A major part of the project is customizing ArcView to create conservation planning, site specific resource analysis, and contract support tools.
Service Center Information Management System (SCIMS)	The SCIMS project improves the service delivery to USDA customers by providing a "core infrastructure" containing common customer, land, and program information.
Conservation Area Resource Analysis and Assessment (CARAA)	The CARAA Project was initiated to improve the process of county-wide resource assessment at the service center by increasing the consistency and scientific credibility of the assessments.
Wetlands and Easements Project	The project will reengineer the certified wetland determination process by developing a Wetland Determination Toolkit that is user-friendly, enables heads-up or Global Positioning System (GPS) digitizing of certified wetlands, and provides customers with a standard package of maps and information.
Geospatial Data Acquisition, Integration and Delivery (Data AID) Project	The Data AID Project was initiated to define and reengineer the business processes associated with the acquisition, integration, and delivery of geospatial data to the service centers for building of the business case for national deployment.
Geographic Information Systems (GIS) Software and Application Training Project	The GIS Training Project was initiated to deliver GIS training to the 9 pilot sites and develop a training strategy on how to deliver GIS training to all the service centers across the nation.
Resource Data Gateway	The Resource Data Gateway is a project to pilot the one-stop-shopping access to geospatial natural resource data.

Feedback from the pilots was used to better understand the geospatial data requirements of the Service Centers as well as refine the processes for integrating and disseminating the data. As a result of this activity, the *Geospatial Data AID National Implementation*

Strategy Plan [A2] was developed. The purpose of this document was to present the recommendations of the Data AID Project Team, document the activities which led to their determination, and establish a framework for the implementation of the recommendations. The recommendations identified in this strategy provide problem statements, case studies, and lessons learned from the project. Processes, standards, testing results, and costs have been used to formulate their recommendations. These recommendations provide specific implementation guidance to those implementing the recommended geospatial data AID plan.

These activities lead up to the current task of developing the Implementation of Geospatial Data Warehouses. Each of the above mentioned documents can be found at http://www.fsa.usda.gov/scdm

In order to make sound implementation decisions one must consider several aspects of the future vision roadmap. One major consideration is the current and near-term state of geospatial and computer technology. Figure 5-1 identifies a few of these technologies and how they may converge to form the future GDW framework.

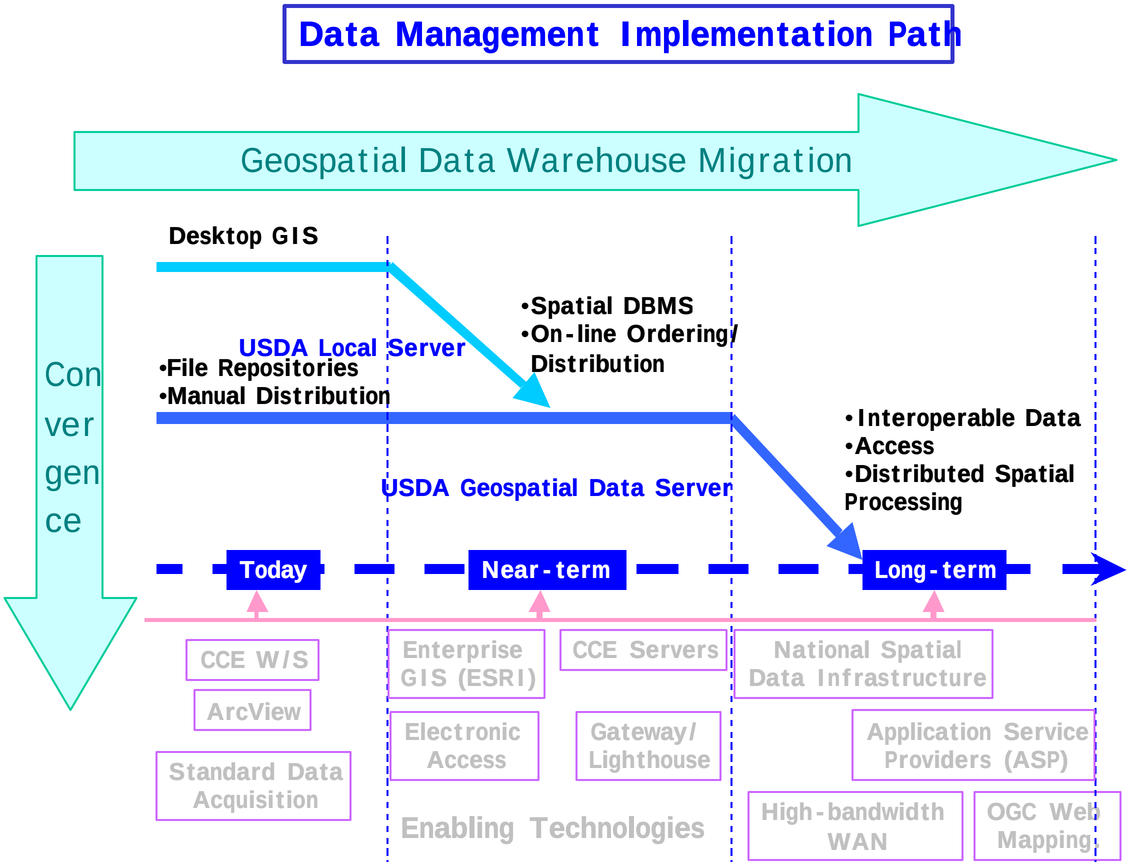


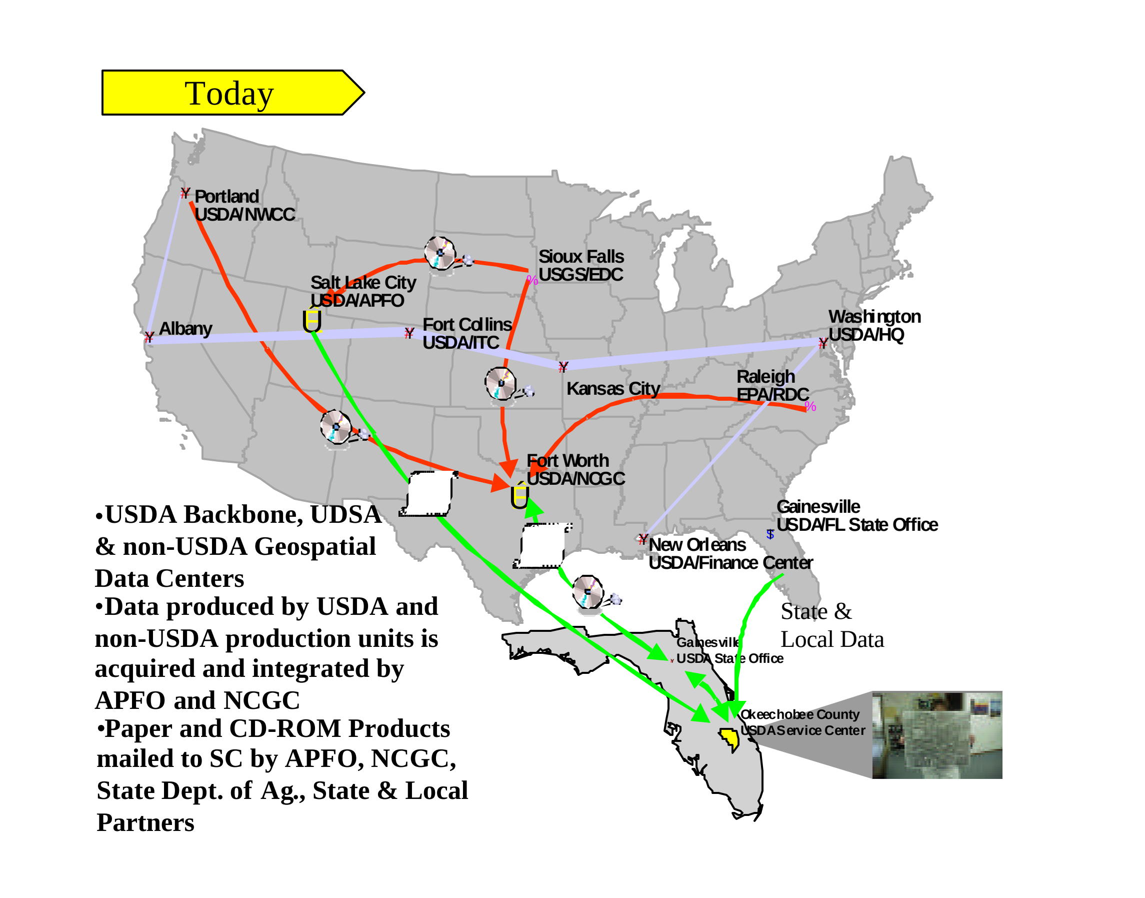
Figure 5-1– Data Management Implementation Road Map

5.1. Today’s Geospatial Data Dissemination Activities

Today’s geospatial data dissemination activities consist of a distributed acquisition, integration and delivery model that was highlighted in the *Geospatial Data Acquisition, Integration and Delivery National Implementation Strategy Plan* [A2]. The document described how the SCA disseminate geospatial data in a widely distributed environment supported by a modest telecommunications infrastructure. As the Common Computing Environment (CCE) Team works to upgrade telecommunications and computers, USDA data production centers work to prepare a common set of integrated geospatial data sets. At the national level, data dissemination is focused on two primary data AID centers located at the Aerial Photography Field Office (APFO) in Salt Lake City, UT and the National Cartography and Geospatial Center (NCGC) located in Ft. Worth TX. The implementation of this dissemination requires APFO and NCGC to acquire geospatial data from other federal agencies and process the data to a level that meets the business requirements of the field office staff. This task is facilitated through partnerships and cooperative agreements with several federal agencies. Once acquired, both USDA owned data sets and non-USDA data sets data are integrated at the county level of geography. Currently, the dissemination responsibility includes organizations at the regional, state and local levels as well.

APFO and NCGC production centers accomplish digital data delivery through a combination of mailing CD-ROM (Compact Disk Read Only Memory) and/or tape and digital download via FTP (File Transfer Protocol). The production centers provide instructions to the SC (Service Center) staff on the proper method to load data on their local server. SCs that have received copies of ArcView®[1], a desktop GIS, can begin to incorporate GIS into their day to day business practices and customer service transactions. Currently, multiple copies of ArcView are distributed through single-user license agreements and loaded onto the local desktops. This dissemination model worked quite well for the nine BPR pilot sites that were established for SC modernization. However, as SCs receive CCE upgrades and geospatial data, a more efficient distribution system must be developed for an enterprise-wide deployment. This vision will begin to be realized during FY (Fiscal Year) 2001 as the elements of the near-term vision are implemented.

[1] ArcView is a registered trademark of Environmental Systems Research Institute, Inc.



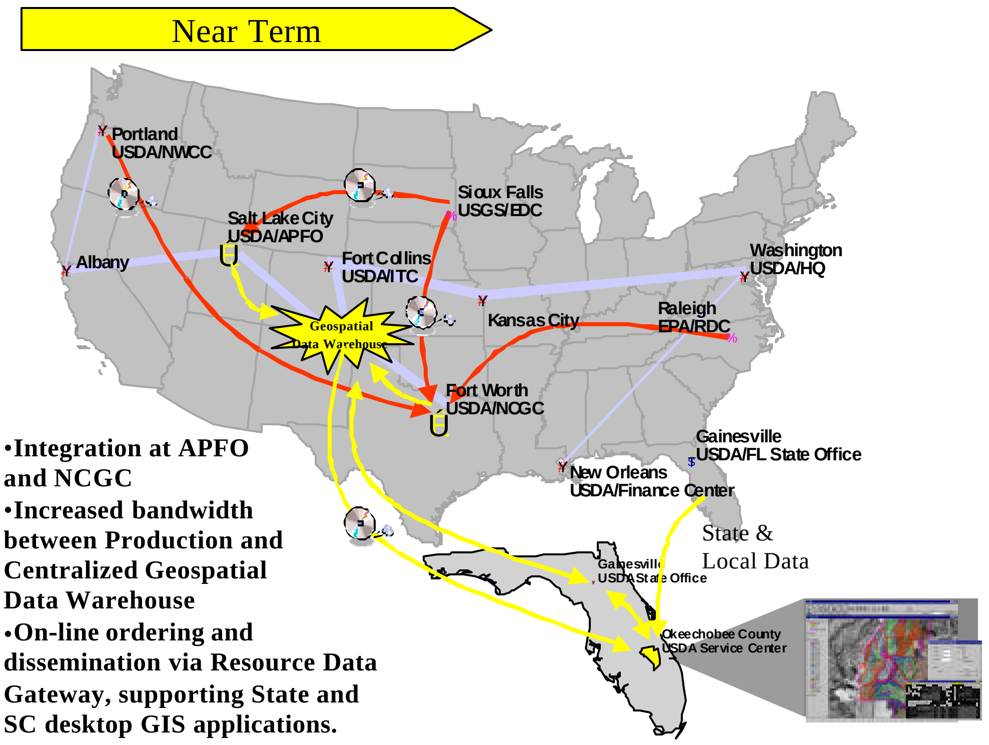
•USDA Backbone, UDSA & non-USDA Geospatial Data Centers
•Data produced by USDA and non-USDA production units is acquired and integrated by APFO and NCGC
•Paper and CD-ROM Products mailed to SC by APFO, NCGC, State Dept. of Ag., State & Local Partners

Figure 5-2 – Today's Spatial Data Distribution Path

5.2. Near-term Geospatial Data Dissemination Framework

The near-term vision of geospatial data access and dissemination is improved significantly by the consolidation (logical or virtual) of geospatial data sets through a unified USDA geospatial portal and the availability of increased bandwidth between the production centers, the USDA backbone and the local SCs. The Electronic Access (EA) initiative plans to bring web farms into production that will enable SCs to utilize the Intranet and Internet to enhance access to applications beyond the SC LAN. However, it is unlikely, in the near term (FY 2001), that all geospatial data will be transmitted via the web for real-time or even one-time data transmission. It is envisioned that the near term GDW will facilitate on-line search, browse and ordering, the automation of CD-ROM ordering and distribution, and piloting data streaming of a few small data layers across the web.

As more applications become available and technological improvements in telecommunications are realized at USDA there will be less emphasis on the storage of geospatial data sets at the local level and more of an emphasis of one or more centrally located data repositories accessed through the USDA Intranet and the Internet. The timeframe for this vision is within the FY 2002-2003 timeframe.

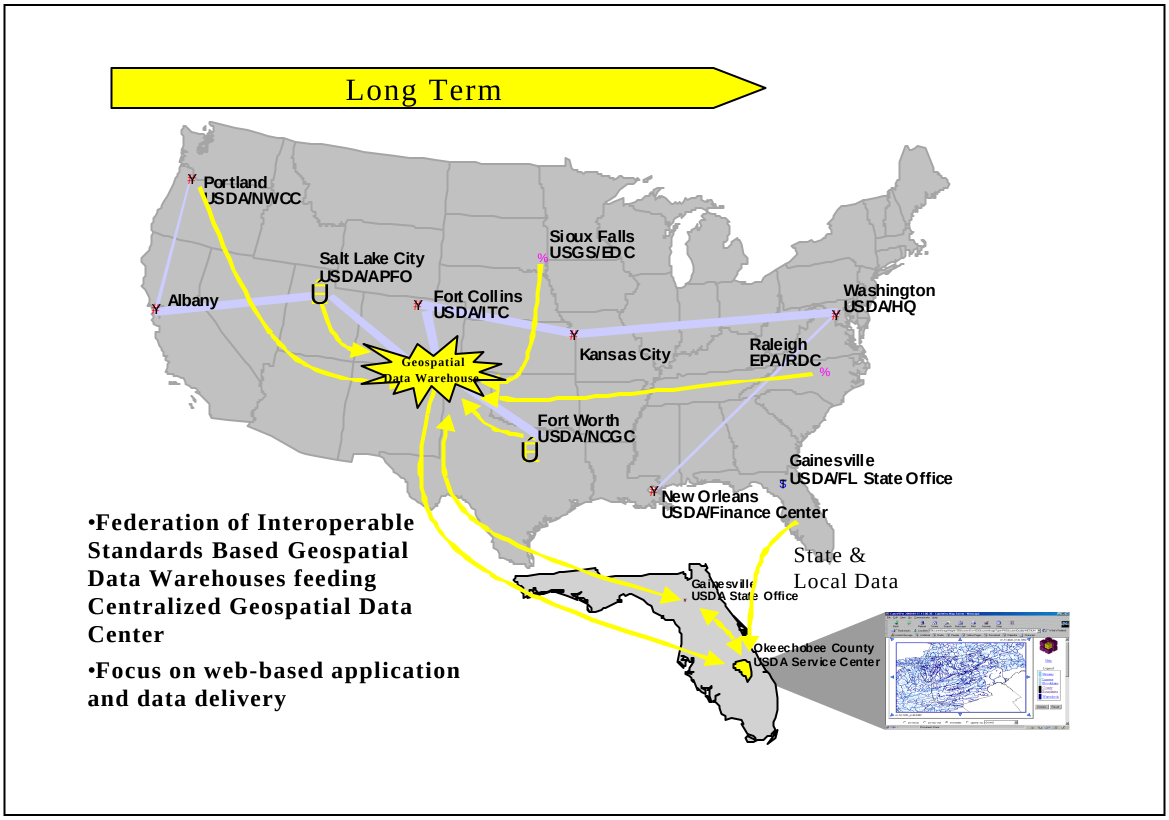


•Integration at APFO and NCGC
•Increased bandwidth between Production and Centralized Geospatial Data Warehouse
•On-line ordering and dissemination via Resource Data Gateway, supporting State and SC desktop GIS applications.

Figure 5-3 – Near-term Spatial Data Distribution Path

5.3. Long-term Geospatial Data Dissemination Framework

The long-term vision of geospatial data access and dissemination at USDA is one that is shared by most federal data providers and many commercial entities. This vision consists of a global network of shared data repositories that conform to mutually accepted open standards, follow inter-operable exchange specifications and utilize common application services. The focus of this vision is to minimize redundant applications and geospatial data sets storage and focus on web-based applications that operate off data stored at central and distributed data warehouses. This framework vision also supports the ability for USDA to concentrate on the dissemination of their owned data sets and have applications access data currently obtained from other federal agencies and partners directly. The benefit of this environment is reduced storage at the local level, access to the most current data available and more efficient and cost effective delivery and integration processes. Additionally, there will be less need to purchase and maintain GIS software on stand-alone desktop environment and more emphasis on applications and services delivered over the Intranet and Internet. However, in order to take advantage of this vision, a high-bandwidth telecommunications infrastructure must be available to support large file transactions and short response times. This vision, although in place to some extent today, will not be fully operational for USDA business until the FY 2003-2004 timeframe.



•Federation of Interoperable Standards Based Geospatial Data Warehouses feeding Centralized Geospatial Data Center

•Focus on web-based application and data delivery

Figure 5-4 – Long-term Spatial Data Distribution Path

5.4. Building a Data Warehouse for the "Vision"

In order to migrate from today's data distribution configuration to the near-term and ultimately approach the long-term vision, a path to data distribution and on-line access must be constructed. The foundation for this architecture is the data warehouse. A discussion of data warehousing is presented in this section to help clarify its use in this plan versus traditional usage of the term. According to Inmon, in *What is a Data Warehouse?*[2], a **data warehouse** is:

- **Subject-Oriented** - oriented around the major subjects of the enterprise. These subject areas can be data oriented (e.g. soils, farms, demographics) or process/function oriented (e.g. conservation planning, crop reporting, lending).
- **Integrated** - data found within the data warehouse has consistent naming conventions, consistent measurement of variables, consistent encoding structures, consistent physical attributes of data, etc. For example, two separate operational

[2] Inmon, W.H. What is a Data Warehouse. Volume 1 No. 1 Prism Solutions, Inc. 1995 TOC

systems may store land area as acres and the other hectares. In the data warehouse, the values would be converted and stored as one or the other.

- **Time-Variant** - All data in the data warehouse is accurate as of some moment in time. For example, the tract of land may represent the ownership boundaries from the time of purchase to the time of sale.
- **Nonvolatile** - There are only two kinds of operations that occur in the data warehouse - the initial loading of data, and the access of data. There is no update of data (in the general sense of update) in the data warehouse as a normal part of processing.

In contrast a **data mart** is a repository of data gathered from operational data and other sources that is designed to serve a particular community of business unit. In scope, the data may be derived from an enterprise-wide database or data warehouse. The emphasis of a data mart is to meet the specific demands of a particular group of users in terms of analysis, content, presentation and ease-of-use. Users of a data mart can expect to have data presented in terms that are familiar. Web enabled data marts and user specific portals are becoming a dominant influence in the presentation of data marts.

In this plan, the term **data warehouse** is used to collectively describe the traditional data warehouse and data mart components. Furthermore, the initial implementation data warehouse described here does, by no means, meet the textbook definition. Rather, this plan lays out the data warehousing technology path for achieving the goals of the *USDA Service Center Geographic Information System (GIS) Strategy* [A1]. The following figures describe a step-wise approach to achieving the Geospatial Data Warehouse vision by building small increments of warehouse over several years.

In the current environment shown in Figure 5-5, the Geospatial Data Warehouse is a loose collection of subject oriented spatial flat files, metadata and tabular data. In most cases, the spatial data is integrated vertically (between **subject** layers), but is loosely **integrated** with attributes from other functional business data. Data integration generally occurs in disparate stand-alone applications on the users' desktop. As for **time-variance** and **volatility**, much of the data does represent a "snapshot" of geographic space in a given time period. However, large portions of the information are completely replaced by newer versions. At present very little geospatial information is stored for historical reference. This may change with time, as previous versions of the soil and CLU data, for example are stored in the warehouses. This environment serves more as a data repository rather than a traditional data warehouse.

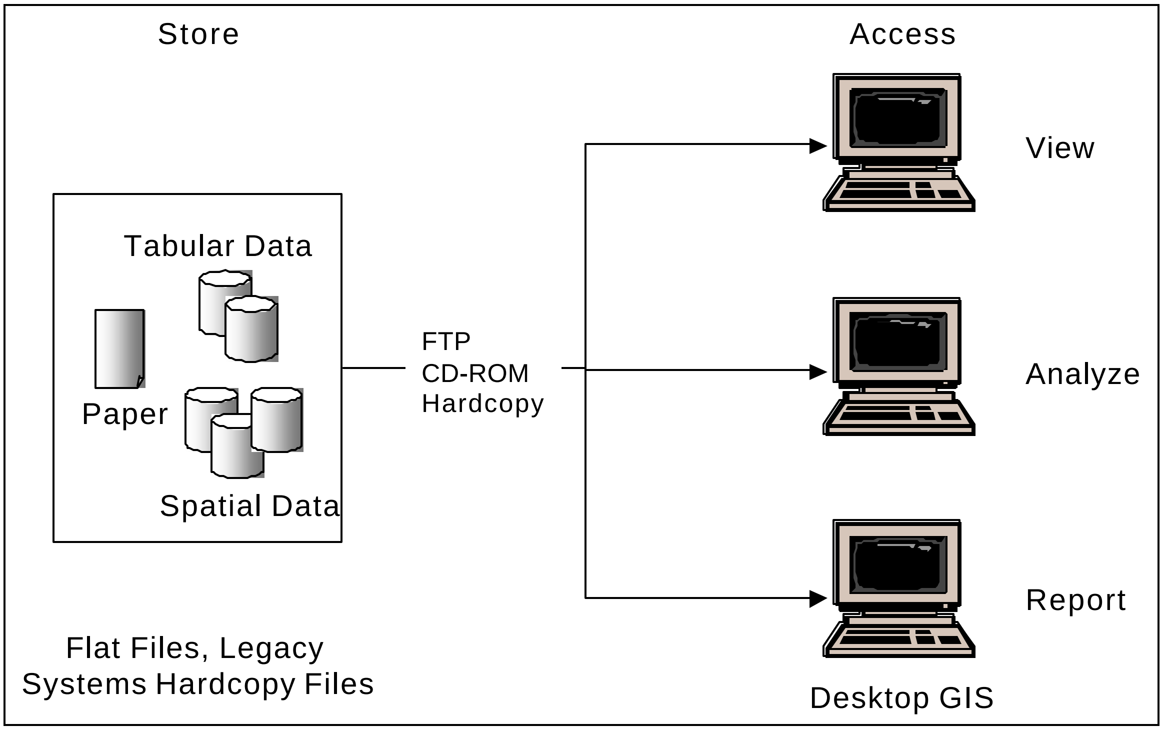


Figure 5-5 – Current Data Warehousing Environment

Enhancements in the near term shown aim to improve spatial data integration by loading disparate databases into one logical, centrally managed geospatial data warehouse. Spatial data is extracted, transformed and loaded (ETL) from operational or production stores to an integrated data structure referred to as the Geospatial Data Repository in Figure 5-6. ETL serves two purposes, it facilitates a tighter integration of spatial data between subjects (vertical integration) and reduces the burden of data delivery to the users by implementing a centralized data distribution operation.

From the user perspective, this improvement simplifies geospatial data discovery, data delivery via CD-ROM and FTP as well as begins to build the framework for on-line delivery of information for viewing, analysis and reporting. Integration of spatial and business data (horizontal integration) are limited to a few cases (e.g. soils data viewer).

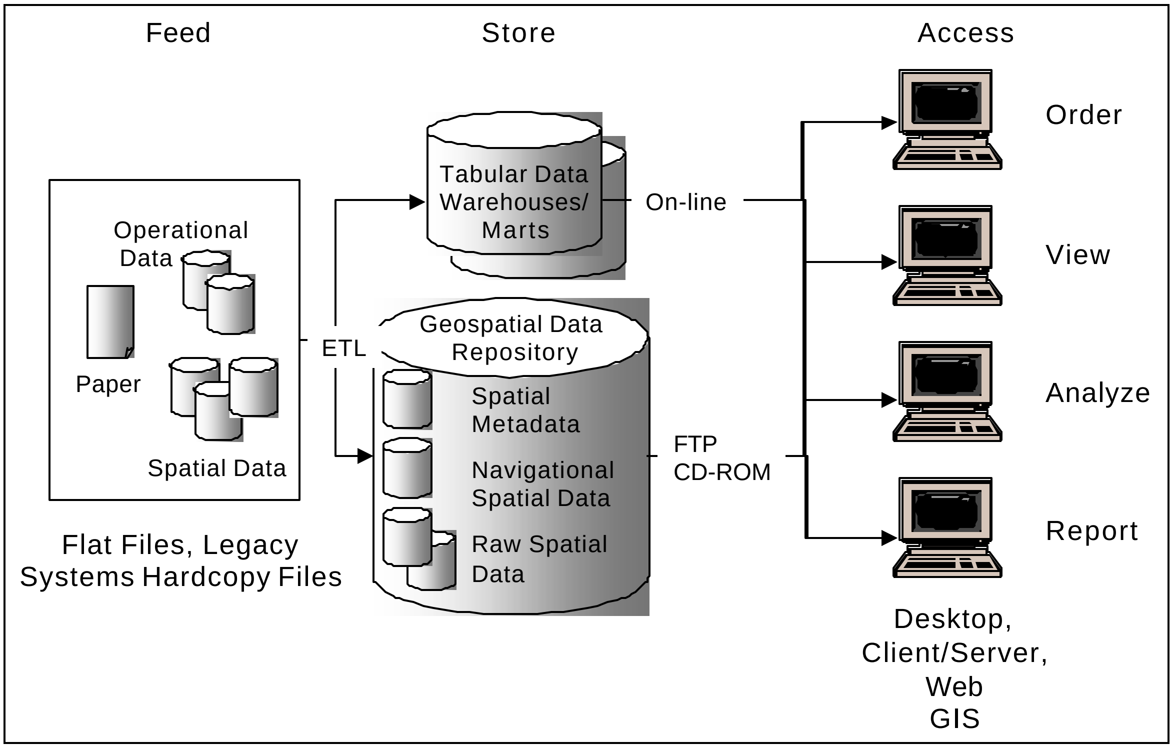

Feed
Store
Access
Operational Data
Paper
Spatial Data
Flat Files, Legacy Systems Hardcopy Files
ETL
Tabular Data Warehouses/ Marts
On-line
Geospatial Data Repository
Spatial Metadata
Navigational Spatial Data
Raw Spatial Data
FTP
CD-ROM
Order
View
Analyze
Report
Desktop, Client/Server, Web GIS

Figure 5-6 – Near-term Data Warehousing Environment

The long-term vision of the Geospatial Data Warehouse illustrated in Figure 5-7 achieves most if not all of the basic goals of data warehousing and data marts.

1. The first action is to load the spatial (data oriented) and tabular (functional oriented) **subject** data into the warehouse. In order to the greatest extent possible, ETL is an automated, rule-based process that continuously feed from a variety of operational and other data feeds.
2. The next step is the vertical (between layers) and horizontal (cross-functional) data **integration**. This requires that an integrated data management organization be in place to handle technical issues such as data modeling, data cleansing and data loading.
3. Integration and implementation of **time-variant** data is the next operation. For spatial data, this requirement will require massive on-line and near on-line storage availability.
4. Finally, the implementation of application specific **data marts** eliminates the **volatility** of the Geospatial Data Warehouse as well as supports the presentation of meaningful information to the broad USDA user community.

This warehousing architecture (data warehousing, data marts) structure will provide new capabilities for the USDA business user that was previously not available, such as decision support and data mining.

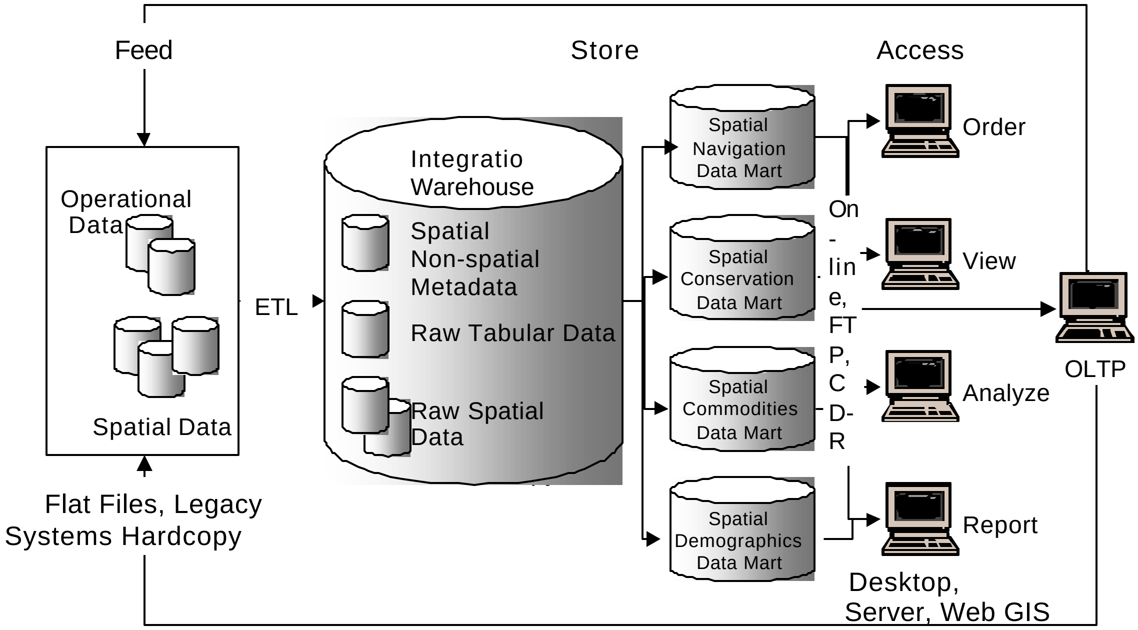


Figure 5-7 Long-term Data Warehousing Vision

6. Roles and Responsibilities

Defining current roles and responsibilities combined with a picture of the "as-is" geospatial data dissemination elements helps clarify interfaces between the various components of the future GDW architecture. The following section defines the organizational and staff level roles and responsibilities as well as mechanisms for managing geospatial data for the SCA.

6.1. Organizational Roles and Responsibilities

Several disparate efforts to manage geospatial data exist within the current Service Center enterprise. Significant progress has been made to synthesize the efforts of these organizations across USDA. However, some redundancy remains. Figure 6-1 illustrates the relationships between each of these organizations. The following section details the major roles and responsibilities of each organization.

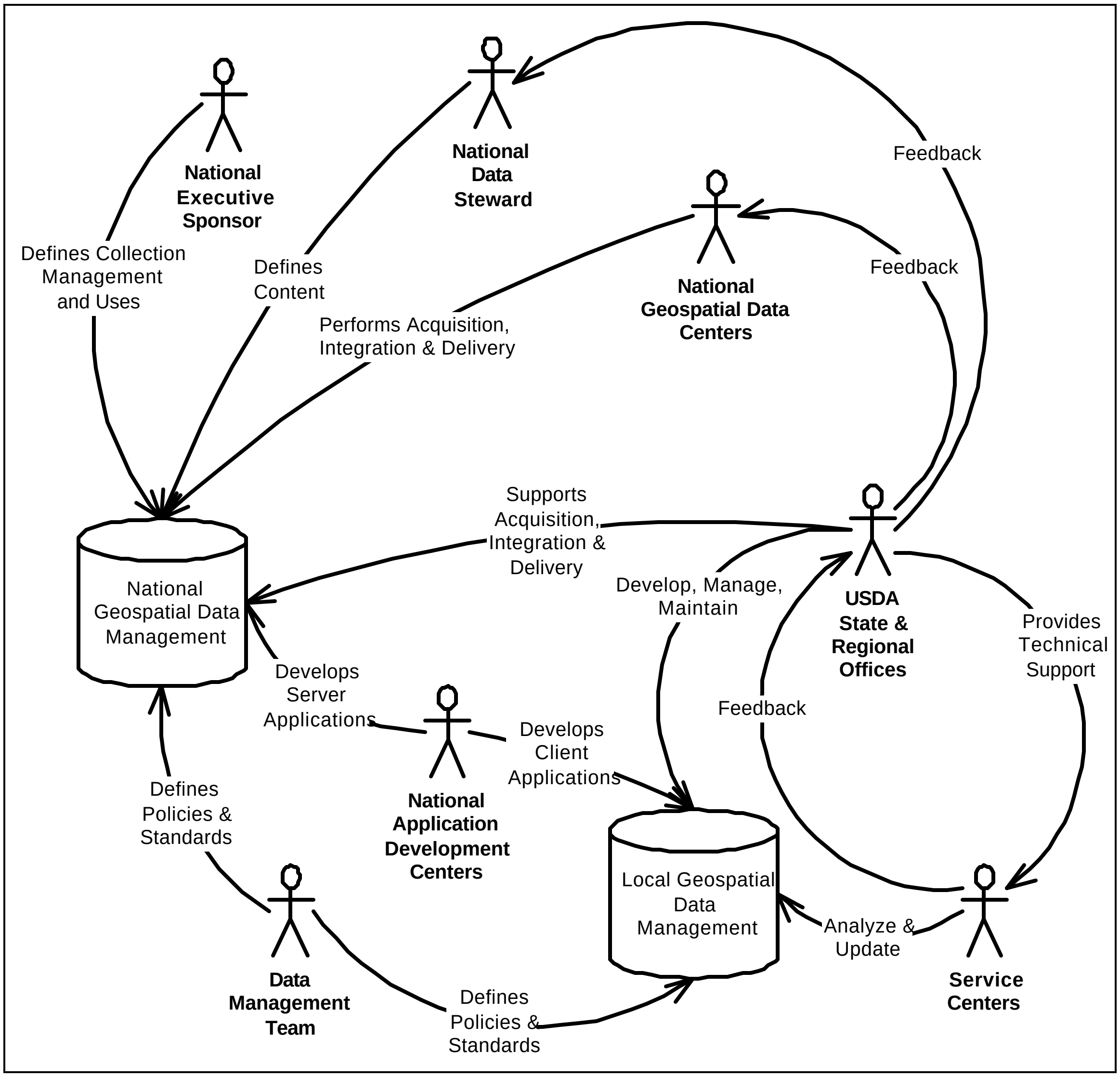

National Executive Sponsor
National Data Steward
National Geospatial Data Centers
Feedback
Feedback
Defines Collection Management and Uses
Defines Content
Performs Acquisition, Integration & Delivery
Supports Acquisition, Integration & Delivery
National Geospatial Data Management
Develop, Manage, Maintain
USDA State & Regional Offices
Provides Technical Support
Develops Server Applications
Feedback
Develops Client Applications
National Application Development Centers
Defines Policies & Standards
Local Geospatial Data Management
Analyze & Update
Data Management Team
Defines Policies & Standards
Service Centers

Figure 6-1 – Major Geospatial Organizational Roles and Responsibilities

6.1.1. Data Management Team

Data management activities and functions for managing geospatial data are distributed in the business/programs and Information Technology organizations within the Natural Resources Conservation Service (NRCS) and the Farm Service Agency (FSA) and to some extent, Rural Development (RD). Coordination activities for managing geospatial data are carried out primarily through the Service Center Data Team (Data Team). The Data Team includes those groups and individuals that are collectively responsible for the effective use, protection, and maintenance of geospatial data assets within the agencies. The Data Team is responsible for coordination of data management activities across all SCA and is acting as an interim interagency team established by the SCI. The Team is responsible for implementing data management principles, policies, standards, and for establishing the overall data architecture.

In support of the enterprise data administration the Data Team will establish a core Data Architecture, to include:

- Develop and maintain data management policies, standards, procedures, and shared utilities and tools for data management.
- Maintain the Enterprise Data Model for all new/reengineered applications among the SCA.
- Coordinate the collection and management of metadata for spatial and tabular data.
- Coordinate implementation of a metadata repository, Computer Aided Software Engineering (CASE), and modeling tools, and other supporting data management software.
- Provide a consolidated voice to the Department and to other government committees on data management issues.
- Establish a framework or structure for data administration processes.
- Facilitate sharing and re-use of common data in Agency and Service Center applications.
- Implement and manage a Change Control function for common and shared data in consultation with the Data Steward and Executive Sponsor.
- Perform Data Administration functions for applications, to include:
  - Resolution of conflicting data names, establishing common lookup tables, setting common domains for sharable data elements, and establishing unique keys and identifiers.
  - Coordinating data administration/management training.
  - Maintain a shared, central metadata repository for use by the Agencies to store and provide access to metadata.
- Establishing standard and sharable data elements to promote data reuse.

6.1.2. National Application Development Centers

Application development centers located in Kansas City, MO (FSA), Ft. Collins, CO (NRCS), St. Louis, MO (RD) and Washington, DC (FSA/Foreign Agricultural Service (FAS) share a combined set of roles and responsibilities that cover geospatial application development and data administration and data management. These roles and responsibilities are described below in generic terms.

- Develops, maintains and supports primary information technology strategic planning and program delivery information systems, databases and applications including the following geospatial applications and databases:
  - Customer Service Toolkit
  - Soil Data Viewer
  - Wetlands and Easements Toolkits
  - Natural Resource Data Gateway
  - Natural Resource Data Warehouse
  - National Soils Information System
  - National Plants Database

  - Ecological Site Information System
  - Integrated Accountability System
  - Service Center Information Management System (SCIMS)
  - Common Land Unit Digitization Tools
  - Land Use
- Develops data models, application architectures and system designs.
- Performs database administration on major agency national database systems.
- Provides National Help Desk support to agency information systems, databases and applications.
- As part of the Service Center Modernization (SCM) Telecommunications Strategy designs, implements and maintains the SCA telecommunications infrastructure.
- As part of the SCM and CCE develops and implements the SCA technical architecture in all agency offices.
- As part of SCM and the Electronic Access Initiative (EAI) designs, acquires and implements server farm capacity at major nodes of the USDA Intranet.
- As part of SCM coordinates with the SCA Interoperability Lab to certify software applications to run on CCE and EA hardware/software configurations.
- Provides technical approval and acquisition support for IT (Information Technology) products and services.

6.1.3. National Geospatial Data Centers

At the national level, two primary data centers exist to acquire, integrate and deliver geospatial data to the USDA Service Center Enterprise. These centers are NCGC located in Ft. Worth, TX and APFO located in Salt Lake City, UT.

6.1.3.1. National Cartography and Geospatial Center

The current data management roles and responsibilities of NCGC are summarized in the bulleted section below. The current production system architecture is summarized in Figure 6-2.

- Coordinate distribution of geodata products and geodata support functions by providing cartography, remote sensing, Global Positioning System (GPS), and geospatial products, services, training, and technical assistance.
- Provide quality assurance for cartography, remote sensing, GPS, and GIS products and geospatial data services, in conformance with NRCS, Federal and industry standards.
- Assist in the development of applications and new technology relating to cartography, remote sensing, GPS, geospatial data and metadata.
- Serve as the NRCS geospatial data clearinghouse to archive and distribute agency geospatial data, to include:
  - Acquire, integrate and deliver geospatial data (USDA and non-USDA) to USDA Service Centers, including associated attributes.
  - Archive geospatial data, copies of soils, Natural Resources Inventory (NRI) and other data at off-site storage locations.

  - Disseminate data by tape, CD-ROM, and FTP on the Internet, according to customer needs and desires.
  - Maintain a toll-free telephone number for customers to order data.
- Provide data stewards for the coordination of graphic layouts, back page contents, and format contracting for reproductions, formatting, metadata and testing for production of CD-ROMs.
- Support the NRI program in areas of data collection, analysis, use, quality assurance and remote sensing.
- Coordinate Agency GPS procurement and equipment use.
- Support national mapping programs like the National Cooperative Soil Survey (NCSS) providing mapping, remote sensing, GPS and geodata assistance.

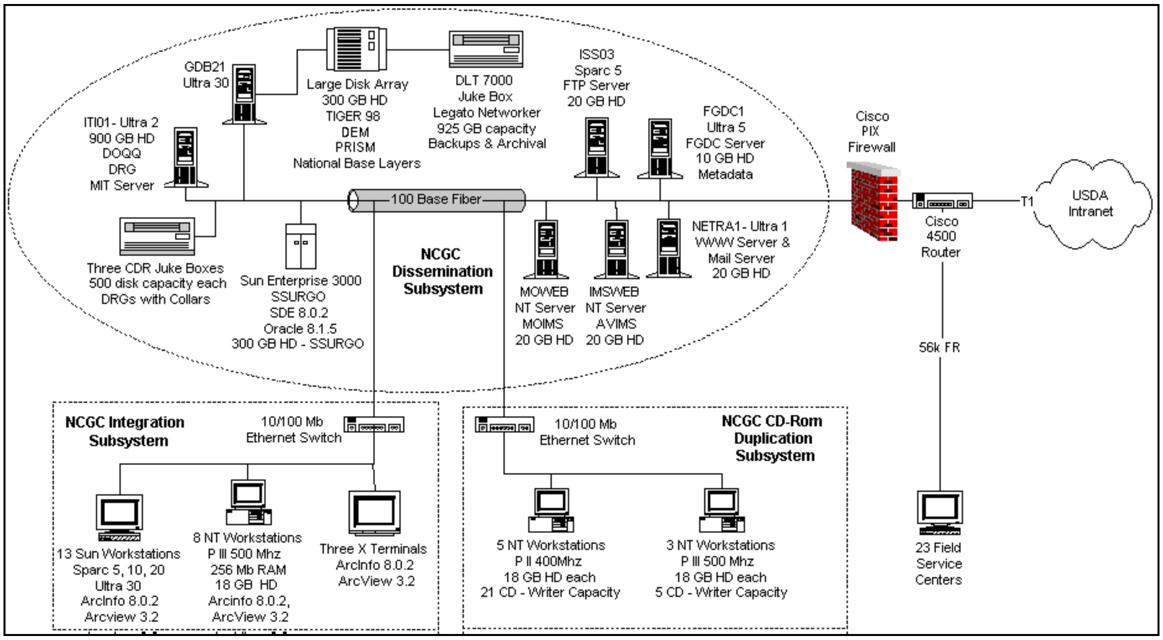


Figure 6-2 – NCGC Geospatial Data Dissemination "as-is"

6.1.3.2. Aerial Photography Field Office

The current data management roles and responsibilities of APFO are summarized in the bulleted section below. The current production system architecture is summarized in Figure 6-3 and the current mosaic production process is detailed in
Figure 6-4.

- Formulate and administer aerial photo/imagery services, to include:
  - Administration of the overall aerial photography and remote sensing programs for FSA.
  - Coordinate aerial photograph/imagery acquisition for USDA.
  - Contract the aerial photo/imagery services for NRCS, United States Forest Service (USFS) and FSA.
  - Conduct USDA aerial photo planning meeting.
  - Serve on the National Aerial Photography Program (NAPP) steering committee.

- – Maintain archive for film acquired by USDA.
- – Provide scale accurate aerial photography for FSA county offices and other customers.
- – Provide quality assurance of NAPP film for FSA requirements.
- – Provide photographic rectification of NAPP aerial photography.
- – Provide photographic enlargements and photo index maps.

- Acquire, integrate and deliver geospatial data (ortho-imagery and Common Land Unit (CLU)) to USDA Service Centers.
- Sell photographic and digital products to other government agencies and the general public.
- Provide technical information and assistance on the use of digital geospatial data and products and services to customers.
- Maintain an archive of geospatial data Digital Orthophotography Quarter Quadrangle (DOQQs) received from the United States Geological Survey (USGS).
- Maintain a metadata repository for digital geospatial and aerial photo holdings.

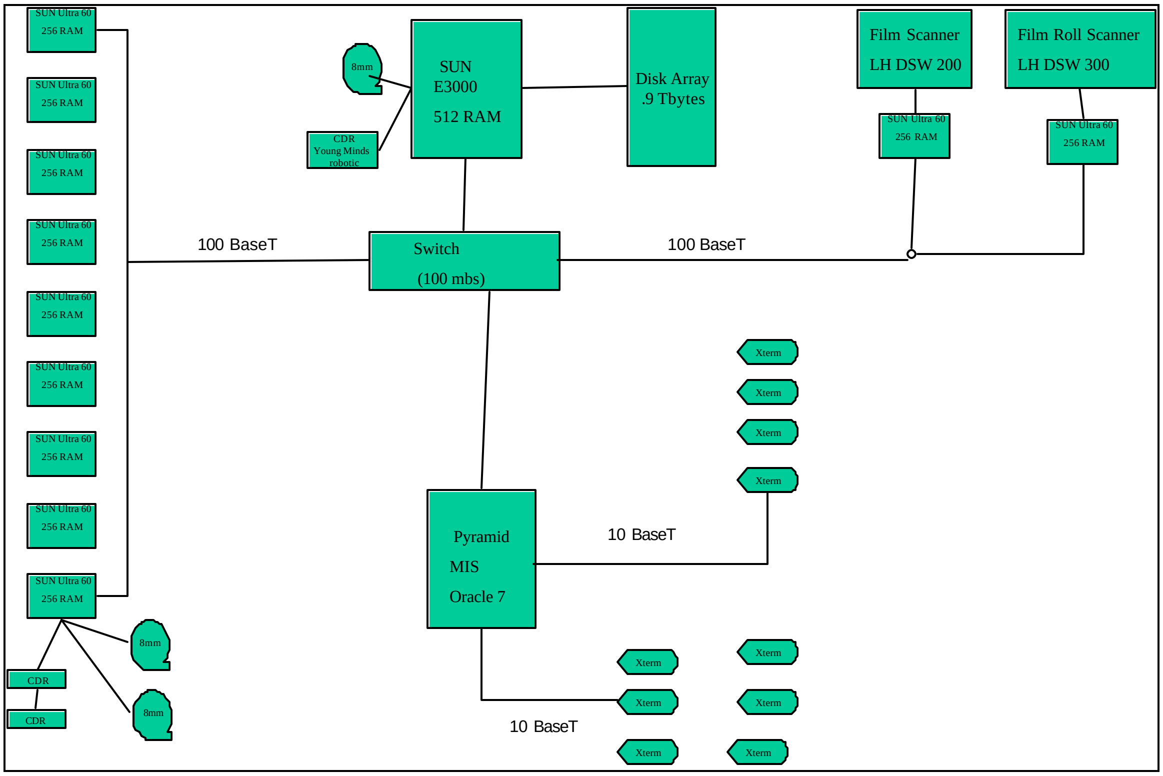


Figure 6-3 – APFO Production System Architecture

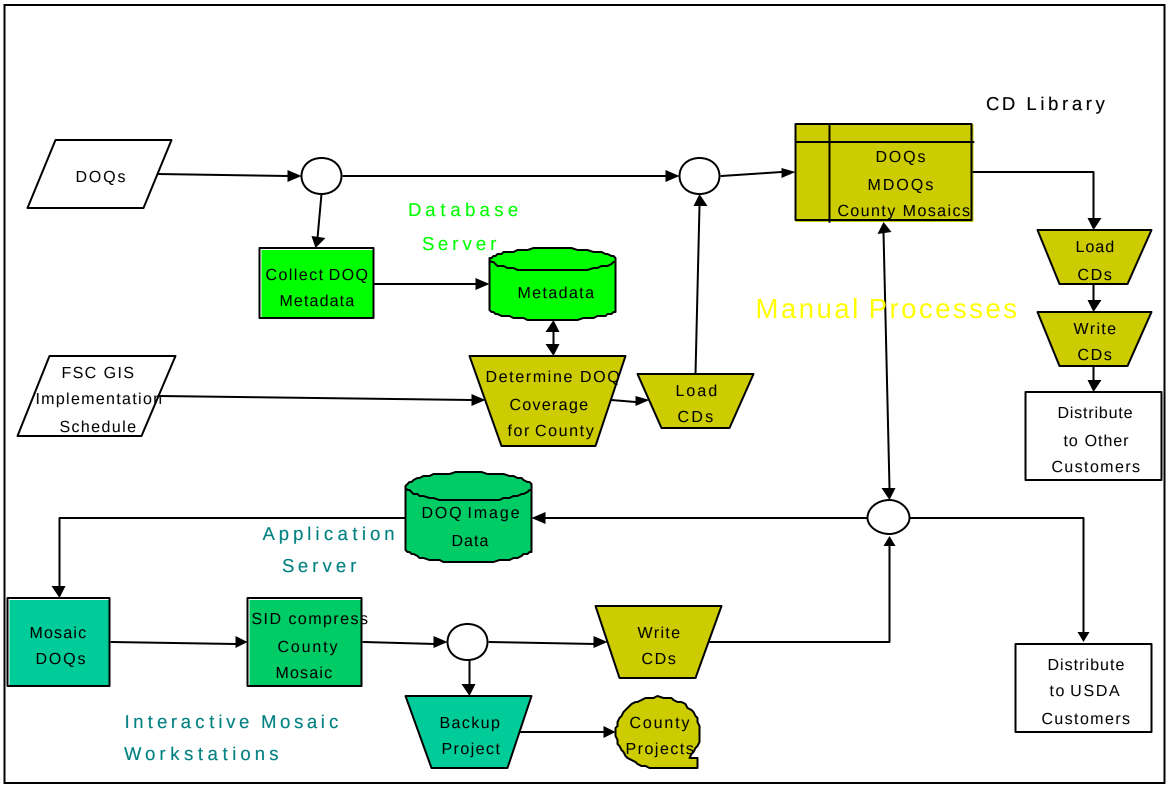


Figure 6-4 - Current Mosaic Production System

6.1.4. USDA State and Regional Offices

The USDA State Offices fall into the category of Geospatial Support and Technology Transfer. This category consists of all persons that direct the development and use of the data assets. They provide support, technical leadership and coordination to the Service Center staff, state and local governments, and outside users of agency information. The USDA State Office GIS Specialist:

- USDA State Offices provide direct support to cartography, remote sensing and geospatial data products, services and technical leadership in support of State and National programs and activities.
- Provide support to field offices including software, hardware, staffing and data for the successful implementation of new geospatial applications.
- Manage and operate computer systems to support cartography, remote sensing and geospatial data products to customers.
- Develop, manage and maintain geospatial data used in the state GIS in conjunction with data stewards and program managers, such as National data layers.
- Assist state staff with quality control by assuring geospatial work performed by state staff, cooperating agencies or contractors conforms to technical standards, policies and procedures and adheres to geospatial data standards.
- Provide user training in geospatial technologies, including:

  - The use of cartography, remote sensing and GIS technologies.
  - The use and protection of GPS receivers.
- Develop and maintain necessary state plans related to advanced technologies, such as a GIS Implementation Plan and aerial photography and ortho-imagery replacement plans.
- Foster communications, cooperative projects and the sharing of geospatial resource data with state or local agencies and other Federal agencies.
- Advise the State Leaders and state staff on geospatial technologies on matters of policy, funding, and personnel requirements.

6.1.5. USDA Service Centers

The USDA Service Centers fall into the category of Data Users. This category consists of all persons who use the data assets, including the Service Center staff, Service Center customers, partner organizations, state and local governments, outside users of agency information, members of the agency business areas, and IT management and staff. The USDA Service Center data user:

- Has responsibility to use geospatial and tabular data, to include:
  - Development, management and maintenance of geospatial data used in county GIS in conjunction with data stewards and state and national program managers.
  - Determine the appropriate use of the geospatial information.
  - Determine the proper definition of data usage.
  - Provide information that supports the extraction and application of data that supports user information needs.
  - Take steps (security, login identification (ID), etc.) necessary to establish access to data stores.
- Provide feedback to application developers and data stewards on the quality, utility, and timeliness of data.

For more detailed information on the roles and responsibilities please consult the *Service Center Data Administration Concept of Operations, August 1998* [A8].

6.2. Staff Roles and Responsibilities

Below the organizational level, the staff provides critical expertise across the organizations to keep the effort running. The following is a list of staff functions that are currently performed and must be performed in the future to maintain a GDW presence.

6.2.1. National Level Operations

The national staff plays a critical role in the sponsorship and stewardship of the geospatial data. Without their high level coordination, the data management operation would become disjointed and dysfunctional. The following sections describe the roles and responsibilities of the National Executive Sponsors and the Data Stewards.

6.2.1.1. National Executive Sponsors

The National Executive Sponsor is a business-area manager who has program responsibility for the data and is accountable for the collection, management, and use of data assets. The person has overall responsibility for the definition of the data, the creation of software systems to collect and process the data, and all issues that deal with data content. In some cases this may be a shared responsibility between several business-area managers from different agencies.

The National Executive Sponsor(s) will:

- Determine data availability, to include:
  - Assessment of existing Agency data collection.
  - Determine if the data is available from other existing sources and coordinate cooperative efforts to obtain the information, as required by *Executive Order 12906, April 1994* [A9].
  - Establish cooperative agreements with non-agency sources of data.
- Coordinate funding for data collection, storage, and maintenance; and for software application development, support, and maintenance. Coordinate with internal and external partner agency management and other disciplines to set development and funding priorities.
- Promulgate and implement the policies and procedures necessary for ongoing data management, to include:
  - The physical data content.
  - The standards for the acquisition and certification of data.
  - The policies for the collection and usage of metadata
  - The procedures for the protection of the physical data assets.
- Designate National Data Steward (s), and other critical data management roles and responsibilities. A data steward is assigned for each national database or sets of data.
- Authorize the release of data and application software to internal and external customers, to include:
  - Certify that software applications meet discipline requirements.
  - Provide guidance and business-discipline support for the development and maintenance of application software necessary for managing the data.
- Have ultimate responsibility for the security of the data assets.
- Manage change as it impacts the business discipline, the needs of customers, and the information delivery technology.

6.2.1.2. National Data Stewards

The National Data Steward is a business-area expert who is assigned responsibility by the National Executive Sponsor for the content of the database. In the case of geospatial data that the agencies do not collect but acquire from other sources, a National Data Steward will be assigned responsibility for the data. In these cases the definition of the data content is usually established and the following roles for establishing these definitions do not apply. The contact point, training, management and help desk duties would apply.

Data steward responsibilities may be delegated to local data stewards who are responsible for portions or copies of a data set. However, responsibility for the definition of the data cannot be delegated. The National Data Steward(s) will:

- Act as the designated authority and point of contact for all business-area decisions concerning the database. Responsibilities include obtaining the needs/requirements from the users, and coordinating with the Data Team on metadata and other data management issues. Also, act as the point of contact for obtaining information on this data and for access to the data.
- Establish and maintain business rules and consistent definitions for data elements, to include:
  - Identification of data domains and relationships.
  - Establishment of data quality and certifications standards associated with the contents of the database.
- Ensure the validity, accuracy, and completeness of the physical data and supporting metadata, to include:
  - Provide guidance for the creation, storage and dissemination of data sets and associated metadata.
  - Certify that the data meets quality standards.
  - Certify that the data is ready for release for internal and/or public use.
  - Implement quality assurance procedures for newly-collected and updated data.
  - Ensure that metadata are collected, approved and certified for release according to adopted industry, Federal and USDA metadata and data management standards.
  - Ensure metadata is made available according to the adopted standards.
- Provide training within the Data Steward's business area, to include:
  - Data management roles and responsibilities.
  - Identification of training needs for data users.
- Provide "help desk" support to the governmental and outside users of data and supporting software.
- Coordinate with agency security officers, to include:
  - Recommend availability, security and access authority for the data.
  - Identify security requirements under the Freedom of Information Act, and for data that must be protected under the Privacy Act.

6.2.2. Geospatial Data Warehouse Operations

The major roles and responsibilities for the successful operation and management of a GDW are well known in the industry and they have been tailored to meet partner agency needs. The amount of effort required for each role will depend upon the diversity, size and activity of the data sets being managed. In some instances a single individual can perform multiple roles, and in other instances more than one individual will be needed to perform a single role. However the tasks to accomplish these roles are essential and must be performed for successful data distribution and maintenance to support Service Center program needs. Each site hosting a GDW component will need to provide support for some or all of these functions.

6.2.2.1. Project Management (Project Manager)

General oversight and responsibility for the delivery of system services in a production environment. Includes the acquisition and managing of resources, coordination with the system's Executive Sponsor, monitoring of production statistics, and general system management responsibilities.

6.2.2.2. Data Stewardship (Business Area Representative)

The Data Steward is a business-area expert who is assigned responsibility for the content of the data. They are the owner of the data and have responsibility and accountability for the actual content of the data and metadata in the system. Responsibilities include the validation, certification and authorized release and dissemination of data, and enforcing rules for maintaining the integrity of the data. During software requirements and design efforts, the Data Steward has responsibility for the defining of all data elements in the application, establishing adequate procedures to ensure the validity of the data, quality assurance on the data model and determining the metadata to be collected to describe the data. The Data Steward may also acquire data from outside sources, or contract for the acquisition of new data.

6.2.2.3. Data Management (Data Manager)

Technical responsibility for day to day management of the data. Works directly with the Data Steward to manage the data including adding new data sets to the database, archiving old data, and quality controls and quality assurance for the data sets.

6.2.2.4. Database Administration (Database Administrator)

The Database Administrator is responsible for the daily administration of the database management software (DBMS, geospatial data engines, etc). The Database Administrator responsibilities include making sure the database is secure and performing as required. They must assure that backup and recovery procedures are in place, monitor the growth of the database and ensure that adequate disk space is available. They monitor performance and take steps to prevent degradation in database performance. They must work closely with the System Administrator to install all database software and patches.

6.2.2.5. System Administration (System Administrator)

This function includes hardware and operating system support. The System Administrator manages daily backup processes, technology refreshes, enforcement of system security, coordinating with communications and other service providers, and general system management and maintenance. Functions include planning and scheduling the installation of new or modified hardware/software, allocating systems resources, managing accounts and resolving hardware/software interface and interoperability problems.

6.2.2.6. Security (Security Officer)

The Security Officer implements and manages Internet and Intranet security procedures, monitors system security breaches and notifies authorities of unauthorized access. They monitor the implementation of security update/patches as needed. Additionally, they ensure the rigorous application of information security/information assurance policies, principles, and practices.

6.2.2.7. Web Administration (Software Engineer)

Responsible for implementation and management of Internet web services. This includes maintaining web servers, web software, telecommunications connectivity, monitoring web site functionality, and integrity, troubleshooting and resolving problems, reviewing, testing, and integrating web pages, collecting and analyzing web site statistics. This position includes Web professionals commonly referred to as webmaster, web specialist, web developer, and web architect.

6.2.2.8. Communications and Network Services (Software Engineer)

This function covers the planning, integration, maintenance, and/or management of networked systems. Functions include maintaining physical network architecture and infrastructure, configuring and optimizing network servers, hubs, routers, and switches, analyzing network workload, monitoring network capacity and performance, diagnosing and resolving problems, making adjustments to ensure proper load balancing, installing, testing, maintaining, and upgrading network operating systems software.

6.2.2.9. Customer Support (Support)

This function includes planning and delivery of customer support services including help desk, troubleshooting, user assistance, and/or training. Functions may include diagnosing and resolving problems in response to customer reported incidents, researching trends and patterns of problems, developing and maintaining a problem-tracking database and coordination and dissemination of data distribution media such as CD-ROM distribution.

7. Geospatial Data Architecture

This section of the document outlines the functional requirements of the proposed geospatial data architecture and presents a logical framework of its components. Also presented, are the evaluation criteria that are used to measure the feasibility of the proposed architecture alternatives.

7.1. Geospatial Data Warehouse Functional Requirements

Part 1 of this plan entitled *Geospatial Data Requirements* defined the data requirements and the high-level use cases for the GDW. This section aims to define high level functional requirements for the GDW. Table 7-1 lists functional requirements derived from the Gateway/Lighthouse Project. This list represents a "first-cut" at defining the USDA Service Center GDW functional requirements. The third column in this table to identifies where each of these requirements would likely be fulfilled in a distributed environment. The term *Central* indicates those functions of the architecture that would be performed at the *web farm* in a distributed environment; *DAI (Data Acquisition and Integration Center)* indicates those components that would be collocated with the data repositories in the distributed environment. In a centralized architecture, all functions are fulfilled by the central site.

Table 7-1 Geospatial Data Warehouse High Level Functional Requirements

Functionality	Description	Location in Distributed Architecture
Application and system operation monitoring	Application monitoring to improve response time to interruption of services.	Central/DAI
Application Web Interface	Non-Outlook interface to data repositories. Real-time data return from client request for data. Functionality defined elsewhere. Soil Data Viewer. Idaho One plan.	Central
Authentication Service	Information control using LDAP and OS authentication. Information that needs authenticated access has not been defined prior to the pilot.	Central
Backup/Restore	System restoration process / disaster recovery activities. Does not include data revision archiving. Return to service target time is one hour. Analysis of the failures that can be reasonable be done in the time is still pending.	DAI
Data Correctness Feedback process	User input to be routed and tracked back to data originator for incorrect data.	DAI
Data Importers	Upload from data providers and sources with conversion as needed. Includes Data Steward controls for allowing/disallowing categories of use for data.	DAI
Data Revision Sequence Checking (Request Manifest)	Assignment of sequence numbers to individual or groups of data to allow checking of "freshness" for download control or update notification.	Central
Data Steward Reports	Generation of what information is stored, controlled or shipped by the Gateway. This may allow for some cache control functionality if request volumes are sufficient. The packets from the Package Builder are the prime data for this volume request tracking.	Central

Functionality	Description	Location in Distributed Architecture
Delivery By CD Generation	Generate request to CDROM operations for shipping with chargeback for costs. Includes e-mail notification.	DAI
Extent Server	Mapping of tagged names to shapes.	DAI/Central
FDGC Metadata Importers	Some metadata is not yet available from the data stewards.	DAI
FGDC Metadata Extraction	SQL extract of data for request. Part of the functional principle guiding information upload with metadata. (XML realization to be included in later phases. Dependent on SQL 2000 XML services.)	DAI
FTP Delivery Service	Including e-mail notification.	DAI
Gateway Catalog Search	Allows a request from the Preview Server, Product Finder, or the Package Server to obtain catalog information.	Central
Metric Collection and Reporting	Processing information collection for scaling and operational data analysis. Use IIS web logs, W2K event logs, Performance Monitor logs, MSMQ statistics, Network Loading statistics, custom statistics	Central
Navigation Server	Includes navigation dialog and display of navigation images for zoom processing.	Central
Package Builder	Provides the directives to the various subsystems for building the information for the requested data. This "dispatcher" will be responsible for the initial time to completion estimate based upon the empirically determined times for various activities. Also provides the compression packaging for the requested data. And notification of completion via e-mail. Determine delivery mechanism to be used.	DAI
Packaging Request Tracking	Service to trace the location and status of any electronically submitted request for data packaging. This allows the end requestor and operations to identify the state and estimated time for a package to be completed. This allows immediate feedback to the user as to the time to package and follow up if e-mail notification is too slow or failed due to erroneous SMTP address.	Central

Functionality	Description	Location in Distributed Architecture
Preview Server	Provide low resolution, high speed image overview for previewing raw geospatial data.	Central
Product Locating Services	Allows users to locate products based on user requirements (place name, area of interest, etc.)	DAI/Central
Raster Clipping Service	Provides the clippings for images from the DRG or Ortho files.	DAI
Request Catalog Construction	Building the themes informational package for requestors. May require different catalogs for different client authorization levels.	Central
Shape file generation	SDE extract and reprojection from database files.	DAI
Streaming Delivery Service	Real-time downloading of a geospatial data products in a variety of formats (shape, geo-tiff, axl, etc.)	Central
Subscription Service	Capture of demographic information from the external uses of the data provided by the Gateway to allow targeted information dissemination (data enhancements, data refreshments, and geographic activity). Microsoft Commerce Information Server for Windows 2000 is not ready for production use in time for pilot.	Central
Tabular Clipping Service	Extraction and subsetting of tabular data (soils data, etc.)	DAI
Tabular Data Extraction Service	Provide a package of tabular data for local or stand-alone use	DAI/Central
USDA Customer Service Toolkit	Functionality has been defined elsewhere. (Summary to be supplied.)	
Vector Clipping Service	Provides the clippings for GIS feature data.	DAI

7.2. Specific Data Requirements for FY 2001

In addition to the data requirements specified in the *Geospatial Data Requirements Document*, the team identified the following specific data requirements for FY 2001. These data requirements are significant because they are one of the determining factors that define the FY 2001 Geospatial Data Architecture.

Table 7-2 illustrates a breakdown of the key layers to be served by the GDW. The table is structured as follows:

- The first column provides some specific details about the size and format of each layer.
- Columns 2-4 specify the requirements for how the data will be served to the end users in FY 2001. On-line browsing means that a view of the geospatial data will be available for on-line (Internet/Intranet) applications, but will not be downloaded as full resolution files to the users local storage. Data streaming means that the data will be physically copied from the GDW to the users local storage or memory for use in an on-line (Internet/Intranet) application. CD and FTP delivery means that geospatial data will be physically copied and sent to the users local storage via CD-ROM /mail or FTP.
- Columns 5 and 6 define whether the associated services are required for USDA and/or public users.

Table 7-2 Specific Data Requirements for FY 2001

Served from GDW	**On-line browsing & backdrop**	**Data streaming**	**CD delivery/FTP**	**Public Users**	**USDA Users**
Ortho (2001) All full resolution, uncompressed ortho imagery in stored on-line in order to produce any of the following formats: Full country TIFF enhanced = 20 TB (on or near on-line) Compressed Ortho MrSID DOQ/MDOQ = 4.6 TB (includes compressed DOQ, MDOQ, compressed county mosaics. ¾ of country complete in SID by end of year.	Pilot JPEG PNG GIF	Pilot TIF SID GeoTIFF	Required CD delivery – minimally, metadata on-line	Desired (all options, including on-line, data streaming and CD delivery) at the cost of reproductio n	Required

Notes for Ortho(2001)

NRCS

- requires all available orthos in MrSID format delivered by CD/FTP and to pilot data streaming of MrSID files
- Full resolution TIFF will be handled by CD/FTP delivery for NRCS

APFO

- required on line storage to support mosaicking
- Create seamless county SID from Ortho TIFF DB
- Create special request SID from Full resolution TIFF stored on-line
- Support on-line web viewing for pilot application development and viewing multiple geospatial data sets

Served from GDW	**On-line browsing & backdrop**	**Data streaming**	**CD delivery/FTP**	**Public Users**	**USDA Users**
SSURGO Soils (2001) 20 GB	Required, accommodated by the Soils Data Viewer	Pilot	Required (NCGC, FTP)	Required	Required
STATSGO Soils (2001) 1 GB	Required	Pilot	Required	Required	Required

Notes for Soils (2001)

- integrate spatial soils with attributes from NASIS and Frozen Soil List, includes interpretations
- pilot to mean that a state or county will test data streaming for the purpose of determining metrics (cost, time, bottlenecks) for national deployment
- Policy of viewing soils and ortho together: only view certified soils that have gone through QC and integrated with ortho
- Approx. 900 SSURGO by end of FY00 and 1200 by end of FY01 (certified)

Served from GDW	**On-line browsing & backdrop**	**Data streaming**	**CD delivery/FTP**	**Public Users**	**USDA Users**
CLU (2001) All information that can be disseminated legally .5 GB	Desired, low priority	Required	Required, Customers may come into the office and get their boundaries on floppy disk	Required	Required

Notes for CLU(2001)

- Does not include SCIMS transactional DB or time series data, only the current state of the data
- Include polygons, farm#, tract#, field#, acres
- Potential partnership with Vantagepoint to give USDA all SID ortho in exchange for CLU
- CLU attributes and customer information added through SCIMS interface for FSC customer use
- Data would need to be replicated up, no access from SC
- CLU is not final until it has gone through the QC process at the SC. Once that process is complete it could be loaded into the warehouse

- Watershed analysis requires access to CLU, need to consider CLU that crosses county boundaries
- Currently in 14 SC and the 11 pilot counties are digitizing CLU
- Jan 01 maintenance will still be handled locally
- 500 counties are to be completed by the end of 01, 5mb/county

Served from GDW	**On-line browsing & backdrop**	**Data streaming**	**CD delivery/FTP**	**Public Users**	**USDA Users**
DRG (2001) TIFF and MrSID 230 GB (TIF) 12 GB (SID)	Required	Pilot	Required	Desired	Required

Notes for DRG (2001)

- Separate out the layers of the DRG, people are doing this for contours and lay them over the ortho
- Use a gap fill layer for orthos
- Consider that USDA is using federal dollars to add value to the DRGs and perhaps the DRGs are required to be public

Served from GDW	**On-line browsing & backdrop**	**Data streaming**	**CD delivery/FTP**	**Public Users**	**USDA Users**
NRI (2001) 1 GB	?	?	Required, Confidentiality agreement applies for spatial component	Required	Required

Notes for NRI (2001)

- Accessed through on-line analysis system for Query and Analysis. It is also used in SAS. Currently no spatial component used in this statistical analysis.
- Goal is to open up NRI to the external customer. Broad query of the on-line analysis system. No feel for number of hits out in the user community.
- NRI points are not released due to restricted nature of data
- Privacy aspect.
- Currently located in Kansas City and Beltsville
- Currently post 17 different tables that address common queries

Served from GDW	On-line browsing & backdrop	Data streaming	CD delivery/FTP	Public Users	USDA Users
Precipitation (2001)	Required	Required	Required	Required	Required

Notes for Climate (2001)

- Drive to integrate climate data from other agencies to make available for broad use. Get everyone to plan together so extracts can be used together. Similar to the OCG model.
- Demand for real time information from water conservation business units.
- Station data and *gridded* data delivered daily to FSA and FAS around country. Weather stations and modeled data from the Air Force.
- Only have precipitation right now, monthly and annual/by state.
- Long term disaster vision for FSA is to serve climate data to the SC's

Served from GDW	On-line browsing & backdrop	Data streaming	CD delivery/FTP	Public Users	USDA Users
Other (2001)	Required	Required	Required	Case-by-case	required

Notes for Other (2001)
Data sets included in the "other" category:

- FSA's Wetland Points
- NWI
- PLSS
- Easements
- DEM/Hydrography
- TIGER
- FEMA Q3
- HUC
- Plants
- USDA administrative areas
- Census
- Geocoded data layers (an example is customer locations for RD guaranteed loan programs)
- Street, county and Zip code data sets are required to perform geocoding

7.3. Geospatial Data Warehouse Framework

Figure 7-1 is a logical component view of the GDW framework proposed for USDA. This view of the architecture provides the lexicon for further decomposition of the architecture. These components can be located in one central location or distributed among several physical CPUs or geographic locations. The arrows between the components represent the communication paths that exist between the components. The general flow through the architecture is represented in the diagram and described briefly below.

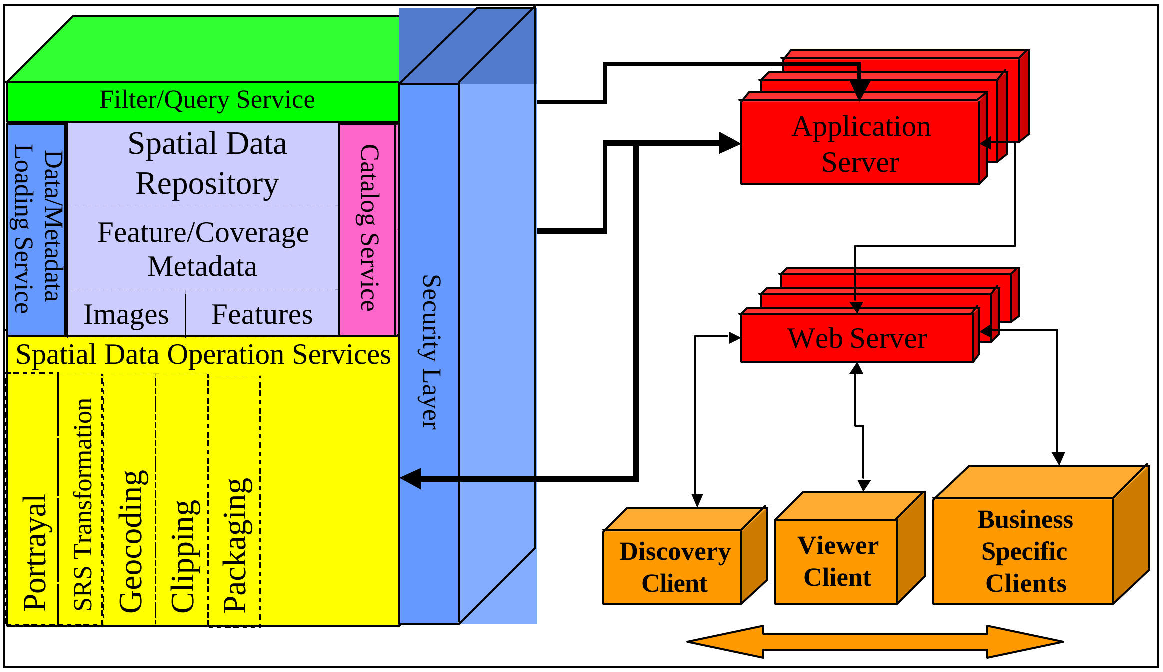


Figure 7-1 – Logical Component-based View of the Geospatial Data Warehouse

7.3.1. Clients

In the component view, clients include both browser based and windows based clients. Client requests are sent across the network to a server via Internet or the USDA Intranet. A client is the requesting program in a client/server relationship. In this framework, clients range from thin, low cost Internet browsers to thick, fully functional business specific applications. In order to support modeling of the architecture, client applications are characterized into three categories; Discovery (thin), Viewer (medium) and Business Specific (thick). These client applications are characterized further below.

7.3.1.1. Discovery Client

A discovery client is used to search spatial data repositories to determine whether or not a particular type of data is available for a particular area. Users discover data by searching and querying on a series of metadata fields and records that contain descriptive information on the data. The discovery client provides direct access to the physical location of the available data. This client is described as "thin" since requests are processed by the web servers and sent back to the client as a response.

7.3.1.2. Viewer Client

A viewer client contains more functionality than the discovery client but does not offer much in terms of analytical capabilities. Examples of viewer clients include Arc Explorer®[3] and HTML (HyperText Markup Language) or DHTML (Dynamic HyperText Markup Language) viewers designed to load map services. These clients often require users to have the latest version of Netscape or Internet Explorer in order to handle the communication requests and responses. Tasks performed in a viewer client are often limited to simple identification of features, zoom and pan functionality and simple attribute queries. Responses to requests by viewer clients are commonly images (GIF, JPEG, PNG), HTML or MIME (Multi-purpose Internet Mail Extensions).

7.3.1.3. Business Specific Client

Business specific clients are custom designed interfaces and applications that have been designed to meet the business requirements of a particular user community. Examples for USDA include the Customer Service Toolkit and the Wetlands application. Business clients are described as "thick" since the bulk of application processing is performed on the client processor. This requires that the client software components reside locally and also may require the user to download data sets to their local desktop in order to process the request. Typical data used by these applications includes binary files (Vector Shape and Raster Mr. SID).

7.3.2. Filter/Query Services

The filter/query services component is responsible for processing data requests initiated by the client browser. Requests may be spatial or tabular in nature or a combination of the two. Filters are used to reduce the list of possible candidate data sets that meet a particular set of criteria. Query services are provided through a query interface, with interaction between the user and the display screen or a combination of the two methods. Interaction may be in the form of pointing to an area of interest or by delineating a user defined area on the display screen. Query results are returned for those records that met the desired criteria.

7.3.3. Spatial Data Repositories

Spatial data repositories are the physical file systems used for storage and retrieval of geospatial data. The repository houses a specific set of geographic feature data, data sets, image data or combination of the three. Metadata for the features, data sets and imagery housed in the repository is also stored in the repository. Data sets are organized logically

[3] ArcExplorer is a registered trademark of Environmental Systems Research Institute, Inc.

in a series of directories and filenames according to the *Standard for Geospatial Dataset File Naming, August 2000* [A6].

7.3.3.1. Feature/Dataset Metadata

Metadata describes how, when, and by whom a particular set of data was collected, and how the data is formatted. Metadata includes attributes such as data name, length, domain of valid values, and definition. Geospatial metadata is simply metadata that describes features and data sets. Data may be a feature, a collection of features (dataset) or an image. Geospatial feature metadata includes such items as a name of the feature, category (i.e., common land unit, soil, hydrology) that the feature belongs to, feature type, stewardship information, and feature attributes and related domain tables. Geospatial data set metadata includes information concerning the content, quality, condition and fitness of use for that particular data set.

7.3.3.2. Images

Images are the electronic equivalent of a hard copy map or aerial photograph. These maps and photographs have been scanned (digitized), processed and georeferenced for use in GIS applications.

7.3.3.3. Features

A feature is a point, line, area (polygon), text, raster or grid in a geospatial data set. A feature includes geometry, topology (if supported), attributes (geospatial and tabular), symbology and labels.

7.3.4. Data/Metadata Loading Service

The loading service is the mechanism used to extract, transform and load data or metadata into the spatial data repository. A loader service could range from a manual process of copying files from portable media to a storage array, to sophisticated database population scripts. In either case the process must be repeatable and have the ability to replace superceded data within the repository.

7.3.5. Catalog Service

The catalog service supports both local and global geospatial data discovery, retrieval of metadata records, browsing, cataloging and indexing of geospatial data. A catalog is essentially a database of information designed to provide information and access to a group of users concerning the availability of geospatial data resources. Each catalog entry contains a resource description. The catalog helps manage the information that promotes data discovery and data access in one comprehensive database. See *The OpenGIS Abstract Specification Topic 13: Catalog Services Version 4, 1999* [A10] for more detailed information.

7.3.6. Security Layer

The security layer of the GDW provides a system to restrict unauthorized access to the data or other services that sit within the security layer. Security will consist of a firewall installed between the application server and the repository and repository services components. In this scenario, a firewall will prevent unauthorized access to the repository components from those outside the USDA Intranet. In an enterprise Intranet environment the firewall blocks domain names and Internet Protocol (IP) addresses originating from unauthorized or threatening sources. If the components of the data architecture implemented for USDA are distributed, additional firewalls may be required. Regardless of the actual physical location, all spatial data repositories and component services will be placed behind a firewall.

7.3.7. The Spatial Data Operation Services

This component provides the functional capabilities that access, process and bundle spatial data for a particular request. These operations convert the data into the appropriate format based requests sent by client browsers. Once data has been processed, it is sent back to the web server through the application server. There are several scenarios for distributing spatial servers. There can be one spatial server running on a single machine, several instances of the same spatial server running concurrently on a single machine or multiple instances of one or more spatial servers running on multiple machines. The spatial data operation services described below are available from the spatial server.

7.3.7.1. Portrayal Services

Portrayal services are the processing steps that get spatial data from the source to the display client. These steps include:

- Data filtering. This operation selects the spatial data to be displayed.
- Data display element generation. This process converts spatial data into a series of display elements used to build a representation of features to be passed on to the rendering service.
- Data render service. This process constructs a map from the series of display elements.
- Display. This service makes the rendered map visible to the user through a viewer client.

7.3.7.2. SRS Transformation

Spatial Reference System (SRS) transformation is a service that converts geospatial coordinates in the data set from one reference system to another. This transformation may also include the transformation between different datums. SRSs are based on a set of standard codes and parameters that are necessary to execute the transformation.

7.3.7.3. Geocoding

Geocoding is a spatial operation service that determines the location of a geospatial feature based on its address. Geocoding can be performed on any geographic feature including an address, intersection, city, state, ZIP code or place.

7.3.7.4. Clipping

The clipping service extracts a set of features, data set or part of an image that falls within the spatial extent of the requested geographic area based on the filter/query service.

7.3.7.5. Packaging

Packaging is the service that delivers the final data to the users according to the parameters of the user request. This includes all of the transformations required by the user as well as the media in which the data is to be delivered (i.e. FTP, tar, zip, CD-ROM, etc.)

7.3.7.6. Application Server

The application server is a background operation that handles the distribution of the incoming requests sent by clients through the web server. The application server determines whether to send requests to the filter/query service, the spatial data repositories via the catalog service or the loading service, or to the spatial data operation services.

7.3.7.7. Web Server

The web server contains the web pages that are accessed by the clients. Transfer of information between the client and the web server is accomplished HTTP (HyperText Transfer Protocol) requests. HTTP is a set of rules used to transfer files on the World Wide Web (WWW). In addition to the web pages, a web server must contain a program that allows the web pages to be served on the WWW, such as Microsoft's Internet Information Server (IIS), Apache or Netscape's FastTrack or Enterprise servers.

7.4. Definition of Evaluation Criteria

Evaluation criteria are used to support scenarios that are economically feasible and likely to succeed based on tangible factors such as organizational acceptance. Evaluation criteria will consider existing efforts, infrastructure, and business requirements. Hardware, software and IT staff supporting current data acquisition, integration and delivery at the data centers represent significant assets that can be applied to the GDW framework. Existing or budgeted resources such as servers, near or on-line data storage systems, media production systems and databases will be counted as assets during the evaluation. These costs include measures that ensure compatibility and integration with existing processes, systems, and business requirements at the data centers. For example, APFO maintains metadata on a range of geospatial data, mainly historical aerial photography. This is not currently part of the delivery package to Field Service Centers, however, APFO is required to make the data available to the general public and recover costs of reproduction.

7.4.1. Cost

One of the key determining factors in the evaluation of proposed architectures is cost. This includes the cost to implement and maintain the system and allow for growth to support the long-term USDA geospatial data dissemination vision. Startup costs include all resources that do not already exist such as hardware and software or that can be upgraded to a newer model or version. Startup costs also include new, contracted, reassigned or re-trained support staff. The startup period is defined as FY 2001-2002.

The second phase of geospatial data dissemination is the implementation/migration phase defined as FY 2001-2004. This phase includes the cost to migrate from the current method of data delivery toward the method defined in the selected scenario.

Finally, the operations and maintenance (O&M) phase of the GDW ensures that the system can remain fully functional at full capacity as well as accommodate growth. Maintenance operations are defined as FY 2001-2006 and beyond.

7.4.1.1. Cost Items

In order to evaluate the cost to implement one of the proposed architectures, the cost of each component must be identified. Any of the GDW architectures presented in this document will include existing resources as well as new acquisitions. Only those resources that are required in addition to existing resources will be identified here. These costs include any new hardware, software and telecommunications necessary to establish and maintain a GDW and the staff and training resources that must be available to operate the GDW.

7.4.1.1.1. Hardware

Hardware costs include the upgrade and/or acquisition of servers, storage, etc. that will be dedicated to one or more of the following functions: data server, application server, web server, FTP server and metadata server. Additional hardware may include X-terminals and workstations for warehouse maintenance and peripherals that support data dissemination including, CD-ROM writers, CD jukeboxes and tape backup units. Additional costs to consider include hardware maintenance.

7.4.1.1.2. Software

Software costs include the acquisition and/or upgrade of software that will be dedicated to the operation of the GDW and associated components. Software includes a relational DBMS such as Oracle or SQL Server to help manage geospatial and tabular data sets on the data server and ESRI's SDE (Spatial Data Engine), a software product used in conjunction with the DBMS to manage the spatial component of geospatial data sets in a relational DBMS environment. Additional costs include software used to operate the services of the GDW and include ArcIMS to manage mapservices and spatial services, Web hosting software, security management software, FTP software and Web monitoring software to manage resources.

7.4.1.1.3. Telecommunications

Telecommunications infrastructure is a critical component for successful implementation of the GDW. An insufficient telecommunications infrastructure will not support the level of access and response times required by USDA, their partners and customers. The Data Team is working very closely with members of the EAI to ensure that the bandwidth planned for FY 2001 will be capable of supporting the near-term vision of the USDA GDW. Additionally, as the GDW expands to meet the goals of the long-term vision, telecommunications will become more critical due the reliance on Web-based data dissemination and Web-based application services. The costs of telecommunications include the upgrade or addition of telecommunications lines between components or centers of the architecture as well as the associated hardware (e.g. routers, switches, etc.)

7.4.1.1.4. Training

Training will be required as existing staff migrates to new roles in support of the GDW. Training will also be required as new technology emerges and as entry-level staff begin to take on more responsibility. Training courses to support the operations of the GDW include Web server maintenance, database management and administration, network architecture and network administration and project management. Training in more specific software products such as Oracle (or other selected DBMS), ArcIMS and SDE will be required.

7.4.1.1.5. Staffing

Staffing costs should cover all costs associated with maintaining a staff to implement and maintain the GDW. In some cases, the staffing requirements may be filled with existing resources. However when those resources are not available, or cannot be retrained, those skills will need to be filled by new or contact staff.

7.4.2. Performance

There are several aspects to performance and as many ways to be measured. In this case, performance is defined as a measure of how well the model works for the client, including access speed, system response time, system availability, data currency and delivery turnaround. Performance modeling will be conducted for the architectures selected by the Data Team as candidates for implementation. The performance measurements will focus on the existing telecommunications and telecommunications that are expected to be in place by FY 2001. In order to succeed, the model must be able to respond to user requests in a timely manner. This response time is ultimately tied to the telecommunications infrastructure limitations. Therefore, the performance modeling will focus on the speed of requests received and the speed of the system responding to a request. The hardware, frequency and size of requests and the responses will be fixed for each of the modeled scenarios. Performance modeling results will be used in conjunction with the other evaluation criteria described in this section and will provide another indicator of whether the proposed architecture meets user expectations and requirements set forth in this document.

7.4.3. Ease of Implementation

The ease of implementation will depend on the existing infrastructure including hardware, software, telecommunications and staffing. Implementation will also rely upon the ease with which new resources can be integrated with existing resources. Outdated hardware, incompatible software, limited telecommunications and a staff that is not flexible to the changing needs of their agency can paralyze the implementation of a model.

7.4.4. Likelihood of Success

The likelihood of success for the selected model will be based on whether the system supports the vision and goals of the three partner agencies and is adaptable within each agency infrastructure. The system will not succeed if the selected architecture cannot be implemented with minimal disruption to existing agency operations.

7.4.5. Supports the Long-term Vision

Success of the model will also be characterized by its ability to grow and adapt as new technologies and as clients and partners adopt new methods. Support will come from agencies that share in USDA’s long-term vision of distributed, interoperable systems.

7.4.6. Effectiveness and Sustainability

The effectiveness and sustainability of the model is based on how well the system meets client requests, the timeliness of implementation and the stability of systems within the context of the Service Center modernization. An effective system must maintain redundancy to protect from system failure, and have a management policy in place to mitigate risk. The sustainability of the system will depend on the ability of the system to be maintained as it grows throughout its lifecycle and respond to changing technology and user demands.

8. Scenarios

To determine the best Geospatial Data Warehouse architecture, several theoretical configurations were proposed as alternatives for the near to long term. These configurations can be grouped into two major groups, which are:

1. **Centralized Data Architecture** - All of the geospatial data is physically located in one central data repository. This means that each of the data layers used by the SCA are physically copied from their point of acquisition and integration to a central data

repository to be co-located with all other layers. This centralized data repository would be the single authoritative source for all USDA geospatial data. The data may be *mirrored* to other machines within a central facility or to multiple facilities, however, mirroring should not be confused with *distributed* data.

2. **Distributed Data Architecture** - Specific layers of geospatial data reside on specific nodes on the network. This means that all of the USDA Service Center geospatial data is never co-located on one node. For example, one data layer (e.g. soil surveys may be located at one facility while another layer (e.g. Mosaicked Digital Orthophoto Quadrangle (MDOQ)) resides at a different facility. Data mirroring may occur at some or all of the nodes on the distributed geospatial network, however, mirrored does not equate to distributed data.

In both cases (i.e. centralized, distributed), the data access interface to USDA geospatial data appears logically seamless to the users of the warehouse.

To facilitate the configuration decision making process, several iterations of both centralized and distributed data warehouse models were constructed. Nine scenarios in all were developed. Within the centralized and distributed categories, some of these scenarios are not vastly different. This was a deliberate attempt to ensure that all technically feasible options were adequately represented, ensuring that political and organizational objectives among the partner agencies and specific business requirements are addressed. These scenarios were then presented to the Data Team for comment and review. Of the nine scenarios drafted, five were centralized and four distributed. The Data Team met in Salt Lake City, UT at APFO on July 31 through August 1, 2000 to review the nine scenarios and reduce the number of candidates based on what could realistically be implemented given the technical (objective) and likelihood of success (subjective) evaluation criteria for the near-term time frame. Two scenarios, one centralized and one distributed were selected from the original nine for further analysis. As part of this analysis, these two scenarios will be modeled, priced and assessed as the baseline for the near-term recommendation of this plan. Table 8-1 summarizes the nine proposed and two candidate alternatives of this plan. The following section presents the rationale for keeping or removing a scenario from the candidate list. More comprehensive descriptions and diagrams are provided for the scenarios selected as candidates for performance modeling.

Table 8-1 - Summary of Nine Geospatial Data Architecture Alternatives

	Scenario Section Number	Scenario Name*	Scenario Description
Centralize	8.1.1	Centralized – mirrored sites, one primary node (Fort Collins web farm or Kansas City web farm), one secondary node	Geospatial data is shipped via network or mail to central location. Data in central location is mirrored to one or more sites.

	8.1.2	Single node (Fort Collins web farm or Kansas City web farm)	Geospatial data is shipped via network or mail to central location (Fort Collins web farm or Kansas City web farm). Data is not replicated to mirror site.
	8.1.3	Mirrored sites – one primary node (APFO or NCGC), one secondary node (APFO or NCGC)	Same as 8.1.1, but centralized at APFO or NCGC and replicated at APFO or NCGC.
	8.1.4	Intra-agency – Farm Service Agencies/Forest Service node	Same as 8.1.2 but strives to gain economies of scale by combining resources with USFS.
	8.1.5	Outsource to external organization (government, academic or private)	Same as 8.1.2 but pays a fee-for-service to a non-USDA entity to host the geospatial data warehouse.
Distributed	8.2.1	National, regional and state store owned data	Current data owners adopt and deploy common USDA Geospatial Data Warehouse distributed framework.
	8.2.2	APFO/NCGC/ Fort Collins web farm/Kansas City web farm (all on USDA backbone	APFO, NCGC, Fort Collins web farm, Kansas City web farm all host USDA Web Farm and disseminate respective data sets.
	8.2.3	APFO/NCGC (not on backbone), Fort Collins web farm or Kansas City web farm (backbone)	Fort Collins or Kansas City utilize existing Electronic Access web farms and hosting single user interface; APFO and NCGC host back-end data repositories and distribution systems.
	8.2.4	OGC/WMT distributed architecture model: internal owned plus external	USDA data repositories are distributed among a system of OGC/WMT enables data servers.

*Highlighted rows selected as candidates for more detailed analysis for FY 2001

8.1. Centralized Data Warehouse Models

The five scenarios presented here focus on the dissemination of geospatial data sets from a central repository. A centralized repository functions as the single authoritative source for all geospatial data sets and performing all the functions outlined in Table 7-1. In order for these scenarios to succeed, the telecommunications bandwidth supporting the GDW must be able to handle the number of requests and size of transactions that are issued. The centralized models presented in this document come in two forms, single site and multiple sites. Single site scenarios operate with only one primary node. Multiple site scenarios operate with a primary node and one or more secondary nodes that are instituted for redundancy, availability and fail over. The secondary node(s) mirror the data sets stored in the primary through site mirroring. Figure 8-1 represents a simple schematic of the centralized model.

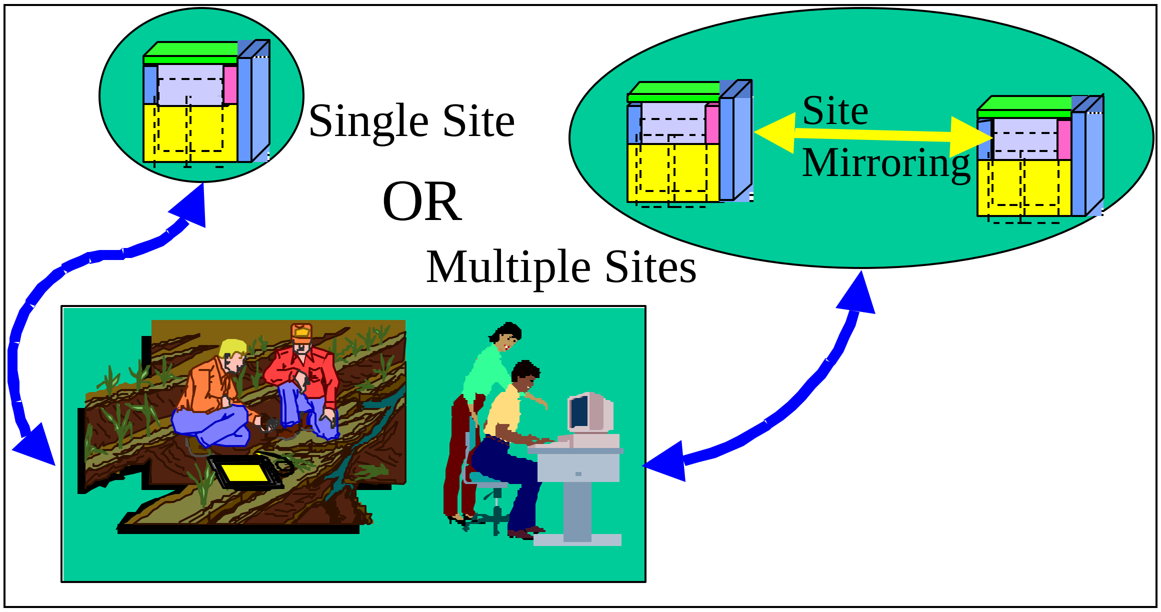


Figure 8-1 - Centralized Warehouse Model

8.1.1. Centralized – mirrored sites, one primary node at the Fort Collins or Kansas City web farm and one secondary node

In this scenario one storage array is designated as the “primary” node and one or more secondary nodes exist. All the components and functions of the GDW in this scenario are centralized to one site, or node located in one facility. Additional support to the primary node is provided by a mirrored GDW located at a secondary node. Data is acquired and integrated at the production centers, as is presently, loaded into the central warehouse and stored on a single storage media array. In this scenario, data is logically centralized, but physically mirrored to multiple network node(s), resulting in redundant storage and access. This secondary mirrored node(s) may be physically co-located or geographically distributed. In this configuration, This is commonly referred to as server “clustering”.

This single node configuration could be implemented at any USDA or out-source location. However the NRCS Web farm located in Ft. Collins, CO or the FSA web farm located in Kansas City, MO would serves as a centralized GDW. The Ft. Collins and Kansas City web farms were selected as candidates to host the primary and secondary node(s) due to their current roles as information technology centers, the availability of major electronic access/web farm infrastructure and their location on the USDA telecommunications backbone providing access to the USDA Intranet and to the public. Access to high-speed bandwidth is a critical component for success in a centralized model. The additional cost of adding bandwidth should be factored in when considering APFO and NCGC as a centralized GDW. This cost should not be used to reject a scenario. At this time RD-STL, located in St. Louis, MO was not considered because RD had not indicated a driving business need for managing geospatial data and the lack of staff with experience in managing geospatial datasets.

8.1.1.1. Initial Evaluation

Based on current knowledge and resources expected for the near-term, including anticipated budget allocations and state of technology, the following positive (pros) and negative (cons) aspects of implementation are summarized below:

Pros:

- Data from the primary node is copied to a secondary or mirror node on a regular basis. The primary benefits of this architecture are to reduce network traffic, ensure better availability and allow the site to arrive more quickly for users topologically close to the mirrored site. A secondary node is an exact replica of the primary site. This architecture facilitates disaster recovery.
- Load balancing between redundant nodes facilitates the availability of data in this architecture. Load balancing divides the total amount of work that a computer has to execute between two or more computers. This allows more work to be accomplished in the same amount of time.
- Provides higher probability of "up-time".

Cons:

- Requires high speed, high bandwidth telecommunications between mirrored nodes.
- Requires redundant system administration staff at "secondary" node. This redundancy essentially causes the cost of implementation to double over the cost to implement a single primary node. The costs are realized in duplicate hardware, software and staffing requirements for the secondary node.
- Requires substantial initial effort to transfer large volumes of image data from data production center to GDW nodes and lesser but still substantial costs for ongoing transfer of image data as it is refreshed.

8.1.1.2. Initial Assessment

This scenario, while technically feasible for FY 2001, is not economically feasible due to the redundant system architecture and staffing required at a secondary node. It is recommended that this option not be a candidate for modeling or implementation.

8.1.2. Centralized - Single node at the Ft. Collins or Kansas City web farm

The single node centralized architecture is similar to the one described in Section 8.1.1 except there is no secondary node mirroring the data stored at the primary node. This scenario is depicted below in Figure 8-2.

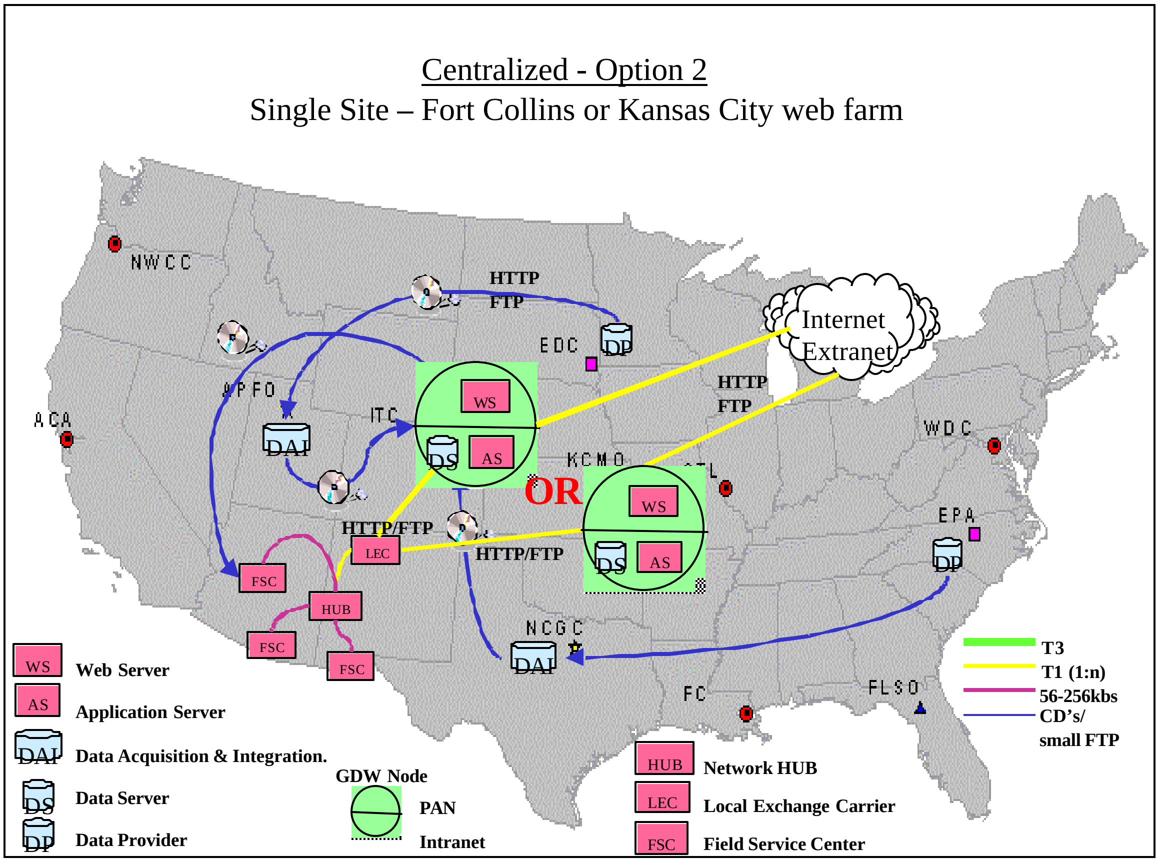


Figure 8-2 - Centralized Option with Single Node

In this scenario the Fort Collins web farm or Kansas City web farm would serve as a centralized GDW. The centralized GDW consists of Web Servers (WS), Application Servers (AS) and Data Servers (DS). Together, these components are represented as the Central GDW Node in the legend of Figure 8-2. The DS within the Central GDW node houses all USDA geospatial data in this scenario. New and updated geospatial data is sent to the GDW from the two Data Acquisition and Integration (DAI) centers located at FSA-AFPO in Salt Lake City, UT and NRCS-NCGC in Fort Worth, TX. In the centralized model all the functions that are identified in Table 7-1 are performed at a central location. In the distributed model these functions are distributed with most performed at the DAIs.

As is currently being prototyped in the Gateway/Lighthouse project, data is delivered through the mail to the GDW at Fort Collins or Kansas City on CD-ROM, tape or through small FTP transactions. Data is requested from the GDW in the form of small HTTP requests through a common Geospatial Data Gateway interface. Browsing requests are fulfilled by relatively small HTTP transmissions over the internet/intranet. Data requests would be fulfilled in a variety of ways depending on the size of the data and the particular database accessed. For example:

- A user may wish to order the full catalog of ortho photography for a given county or field office service area, but does not have the bandwidth or time to download that data on-line. In that case, the user may wish to order a set of CD-ROMs be created, packaged and sent via the mail. The location of the creation and packaging process could be one or both of the Data Acquisition and Integration Centers or the GDW site. That decision is still under consideration. In either case, only the CD production system is significantly impacted.

- A user may wish to download one soil survey for the entire county. Depending on the level of service of the customers' connection, the soil survey may not be large enough to warrant a CD-ROM delivery, but is too large to utilize on-line. In this scenario, a zipped file would be packaged and staged for subsequent FTP downloads. This could be at the time the user places the order or at a later time after the file has been staged and the user notified. The location of the staging area in this scenario is co-located with the GDW node. As increases in bandwidth are implemented the ability to utilize datasets on-line and send larger packets of data will replace CD-ROM delivery.
- A user may wish to simple use geographic data stored in the GDW as a backdrop or to perform simple analysis on a web application. In this case an image (e.g. GIF, JPG) of the data or vector data is streamed to the user's client application in real-time. The user may choose to retain a copy of this view for later use. This scenario is considered only for prototyping in the FY 2001 timeframe. It is assumed that constraints will not afford this solution in the prototype.

8.1.2.1. Initial Evaluation

Pros:

- Consolidates data management into one physical site (staffing, hardware, software, and telecommunications).
- Allows dissemination element to be located close to a single high-speed dissemination hub.
- Eliminates the need for high-speed connections between distributed components of the GDW.
- Eliminates the need for multi-node synchronization.

Cons:

- Current, data management, data stewardship and data ownership are geographically and organizationally dispersed. Departing from this status quo could have a high impact on specific agency business requirements.
- Data must be transmitted via portable media from the data producer to the GDW. This method is currently used to move data from APFO and NCGC to Fort Collins to populate the RDG pilot. This is a labor-intensive process that adds additional costs such as media. Additionally, problems reading the media exist on the Fort Collins side.
- Substantial initial effort to transfer large volumes of image data from data production centers to GDW nodes and a lesser but still substantial cost to transfer image data as it is refreshed.
- Creates lag time between data production and data dissemination.

- Creates single point of failure (no fail-over node). However, this risk is minimized by off-site storage and high availability hardware.

8.1.2.2. Initial Assessment

This scenario does not incur the additional cost of maintaining a secondary node(s) as required in the scenario described in Section 8.1.1 and should be considered as a candidate for modeling. This scenario has a relatively low implementation cost since much of the architecture infrastructure already exists at each of the proposed data centers. Significant issues concerning data ownership, remote data administration and data packaging remain outstanding.

8.1.3. Centralized – mirrored sites, one primary node at APFO or NCGC, one secondary

The architecture in this scenario is identical to that described in Section 8.1.1 except that the primary and secondary nodes are located at the current DAIs, APFO and NCGC, instead of one of the Electronic Access Web Farms, Fort Collins or Kansas City.

8.1.3.1. Initial Evaluation

Pros:

- This architecture facilitates disaster recovery.
- Load balancing between redundant nodes facilitates the availability of data in this architecture.
- Provides higher probability of "up-time".
- Locates data closer to data stewards and data managers.
- Data management and data packaging are at same site.
- A high percentage of infrastructure for primary and secondary nodes already exists.
- Due to existing business requirements, a high percentage of primary and secondary node staffs already exists.

Cons:

- Requires high speed, high bandwidth telecommunications between mirrored nodes. This infrastructure is not currently planned during the FY 2001 EAI.
- Some redundancy of system administration staff between "primary" and "secondary" nodes. A percentage of staff required for secondary nodes already exists.
- Separates the application servers from the data servers at the Fort Collins web farm or the Kansas City web farm.

8.1.3.2. Initial Assessment

This architecture is identical to the one described in Section 8.1.1 with the exception of the node locations. Therefore, due to economic feasibility, it is not recommended for modeling or implementation for FY 2001. This should be considered as a configuration alternative or growth path for future years.

8.1.4. Centralized – Intra-agency, Farm Service Agencies/Forest Service node

This scenario is identical to the single node scenario presented in Section 8.1.2 except the single node is the shared responsibility of all three SCA plus the US Forest Service. The node would host shared data assets as well as those data sets that are unique to each agency. The single node location could be situated at the shared facility presently occupied by both APFO and USFS Geographic Information and Technology Center in Salt Lake City, UT or some other mutually agreed upon location. In addition to the EA Web farms that receive requests from USDA clients, the Forest Service would have their own Web farms that would handle requests from their users.

Preliminary discussions with USFS indicate that there is interest in a shared data repository.

8.1.4.1. Initial Evaluation

Pros:

- Minimize data management costs by combining architecture resources of agencies within UDSA that have similar data storage and delivery requirements.
- Agencies that utilize common data sets that conform to similar standards (or can easily be transformed, such as DOQs) eliminate redundant storage and management of duplicate information.

Cons:

- Shared data resources may expose sensitive data sets to non-farm agency personnel. This may or may not be an issue depending on whether or not the dataset contains any sensitive customer or producer information. Proper security measures should prevent any security breach.
- Data owners may not be physically located with their data sets. Data stewards must trust the hosting agency to ensure availability and access to their assets.
- Conflicting or competing requirements, funding limitations and different policies may impede the overall progress of each agency.

8.1.4.2. Initial Assessment

The USFS and USDA share common goals in terms of geospatial data standardization and dissemination. The experiences of these two agencies can be combined and used to move towards establishing a collocated data repository that could serve both agencies needs. Although not a likely option for FY 2001, this scenario should be pursued for future implementation whether it is implemented as a single repository for all USDA and USFS data sets in a centralized scenario or as another node on a network of distributed repositories.

USDA and USFS should work together to determine where commonalties exist in each agency's core data sets. A comparison of business requirements would help to determine data sets that are redundant and can be universally shared. Additionally, USDA and USFS need to compare their data set standards to ensure that standardization is not compromised in a shared repository scenario. Smaller efficiencies at APFO and other co-located sites should be examined such as combined CD production and shared telecommunications costs.

8.1.5. Centralized – outsource to external organization (government, academic or private)

Again, this scenario is identical to the single node scenario presented in Section 8.1.2 except the single node is established and maintained by an external organization. This organization could be another government agency (e.g. USGS), an academic institution or a private sector company. All hardware, software, telecommunications and the staffing are the responsibility of the outsourcing organization. USDA would need to provide the organization with the most recent versions of the data sets and ensure that the USDA user community had adequate bandwidth to access the node.

One agency exploring the outsourcing model is USGS. USGS recently submitted a request for information (RFI) seeking alternatives to their current, contractor based, data and information dissemination center at their Sioux Falls, SD Earth Resource Observation System (EROS) Data Center (EDC). This RFI requested interested parties to respond to how they would handle the distribution of data as a service to USGS. The primary data types are quadrangle based Digital Raster Graphs (DRG), Digital Elevation Models (DEM), Digital Line Graphs (DLG) and Digital Ortho Quadrangles (DOQ). The volume of DOQ alone is estimated to be on the order of 11-12 terabytes.

Respondents were asked to comment on data volume, locating data through catalog searches, organization of data, Internet delivery dependencies, CD-ROM delivery plans, pricing structure and cost recovery, format conversion services, experience and why the respondent would want to undertake this type of service. The results of the RFI indicate that few organizations are interested in providing this type of fee-for-service operation. Twenty-four responses were received and reviewed. Most respondents offered software-based solutions to enhance dissemination instead of a service to replace the current operation at USGS.

USGS plans to use the information gathered through the RFI process to generate several different scenarios on how best to proceed. The next logical step for USGS would be to release a request for proposal (RFP). An RFP is more likely to generate interest in the commercial world than an RFI. The fact that an RFI was released instead of an RFP may have reduced the number of responses possibly limiting the number of service providers that would be interested in this type of opportunity.

USGS experience with the RFI and how they decide to proceed is directly relevant to the USDA outsourcing scenario presented here and should be monitored closely. One area that is of particular interest to USDA is a data maintenance subscription service that would make sure customers are aware of updates once data has been shipped and/or automatically ship updates. USGS has plans to make this service available by the end of calendar year 2000 using their Earth Explorer system. This system will provide notification or delivery of updates.

8.1.5.1. Initial Evaluation

Pros:

- Potential cost savings in terms of hardware, software, telecommunications and staffing requirements. All costs are included in startup and maintenance costs of the service.
- Follows Federal initiatives to outsource IT services to non-government entities.
- Allows USDA to focus on USDA business, not IT.
- As an ongoing operating expense, budget needs might be easier to plan and justify.

Cons:

- Data managers may not have administrative access to their data sets.
- Service level agreements may not provide adequate information security.
- Requires funding level commitments that could cause loss of agency FTEs (Full Time Equivalents).

8.1.5.2. Initial Assessment

This option should remain under consideration, and continued to be pursued for the near term, especially for non-USDA data. USDA should continue to work with USGS to learn and benefit from their experience in outsourcing alternatives. Cost information should be obtained and compared with the cost to implement in-house dissemination.

8.2. Distributed Warehouse Models

The four distributed scenarios presented here focus on dissemination of geospatial data sets from multiple locations. A network of distributed warehouses potentially including USDA, non-USDA, local, state, regional and national nodes as hosts the geospatial data sets on-line. The distributed nature of this model should appear seamless to the users of the system. The only impact to the user should be in performance response time constrained by slow or disabled nodes on the network. Distributed architectures require each node to support adequate telecommunications and possess the required components of the GDW to support data dissemination from their site. Figure 8-3 represents a simple schematic of the distributed model.

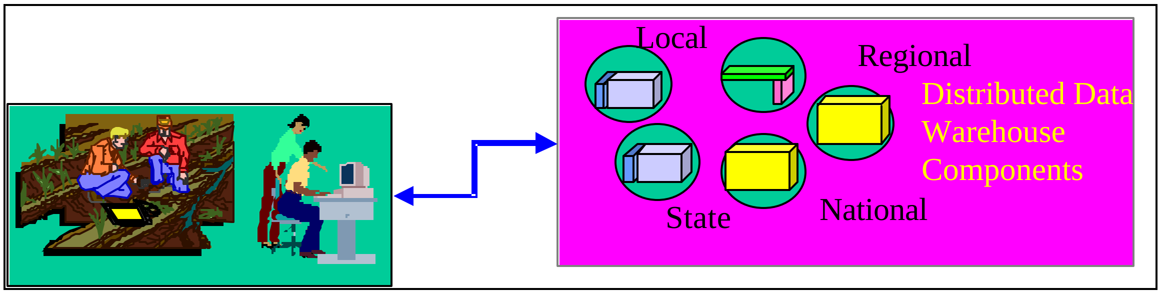


Figure 8-3 – Distributed Warehouse Model

8.2.1. Distributed – National, regional and state store owned data

This distributed data model is basically an enhancement of the current model, where data dissemination (internal USDA and to the public) occurs at many levels of the SCA. APFO and NCGC still function as DAIs, receiving data from external USDA data partners and providers. UDSA agencies that are currently responsible for the production and maintenance of their own data sets, such as the National Weather and Climate Center, would not necessarily send their data to one of the DAIs for integration and dissemination. Rather, each organization would have the option to be responsible for ensuring that their data is available and conforms to the standards established by the Data Team. In order to implement certain distributed models some distribution nodes may need to acquire additional software, hardware, telecommunications and staffing in order to meet availability requirements.

8.2.1.1. Initial Evaluation

Pros:

- Data sets are stored and maintained by the data stewards at the point of production, providing one authoritative source.
- Eliminates the cost for USDA to host a single centralized node.
- Promotes extensible framework.
- Distribute the cost of the GDW to other national and non-national entities.

Cons:

- Increased cost of data management components (staffing, hardware, software, and telecommunications). Not all "nodes" may be able to afford the cost to disseminate their data by FY 2001, causing users to receive inconsistent service until the distribution node could meet delivery requirements. It is more realistic for nodes to become available as technology and bandwidth advances are more accessible.
- If sufficient bandwidth is not available response time may be impaired due to the number of data stores that may need to be visited per request.
- Each distribution node becomes a single point of failure.
- Security procedures need to be maintained, may become complicated.
- Difficult to ensure that external organizations provide data meeting SCI standards for content and format.

8.2.1.2. Initial Assessment

This scenario is a very desirable option and it conforms to the long-term vision of USDA data dissemination, however it does not conform to the vision for FY 2001 in terms of technical and economic feasibility. This architecture is too costly for implementation at this time due to the limited telecommunications infrastructure that currently exists between all the potential nodes. Additionally, nodes would need to bear the additional cost of becoming a data center which many state and local operations are not prepared to support at this time. The recommendation for this scenario is to maintain contact with regional, state and federal agencies in terms of their implementation of hardware, software and telecommunications infrastructure and maintain a dialog in terms of data partnerships and data sharing.

8.2.2. Distributed – APFO/NCGC/Fort Collins Web Farm/Kansas City Web Farm (all on USDA backbone)

In this scenario, the existing USDA telecommunications backbone is expanded to include high bandwidth availability to APFO and to NCGC. APFO and NCGC function as GDWs in addition to their current role as DAIs. These two production centers would need to acquire the additional hardware, software, telecommunications and staffing that would bring them up to GDW status. Additional capabilities would be added to the Fort Collins and Kansas City web farms to bring them up to par with the requirements of a GDW as opposed to their role in EA as Web farms only. Dissemination responsibilities in this scenario are shared between all four centers. Each center would house and maintain their own data assets and would be responsible for making sure backup and recovery mechanisms are in place. A variation on this scenario is to connect the two DAIs directly to UUNet. This variation does not alter the needs of the two DAIs since they still require increased telecommunications.

8.2.2.1. Initial Evaluation

Pros:

- Data sets are stored and maintained by the data stewards at the point of production, providing one authoritative source.
- Distributes the cost for USDA to host a single centralized node.
- Promotes extensible framework.

Cons:

- Increased cost of data management components (staffing, hardware, software, and telecommunications).
- If sufficient bandwidth is not available response time may be impaired due to the number of data stores that may need to be visited per request.
- Each distribution node becomes a single point of failure.

8.2.2.2. Initial Assessment

The implementation of this scenario would require a significant investment in the telecommunications infrastructure between APFO and NCGC to Fort Collins or Kansas City web farms in addition to the cost to upgrade the hardware and software infrastructure in both locations. The feasibility of upgrading APFO and NCGC to Web Farms based on FY 2001 budgets is possible. A more realistic approach to this scenario is for nodes to become available as technology and bandwidth advances are more accessible.

8.2.3. Distributed – APFO/NCGC (not on backbone), Fort Collins/Kansas City Web Farms (backbone)

This scenario most closely resembles the current model of data dissemination. Fort Collins and Kansas City web farms are now part of the dissemination network and function as GDWs in addition to their role as EA Web farms. Fort Collins web farm in this scenario hosts the NASIS and Plants databases and Kansas City hosts the national level CLU data set. This scenario is detailed below in
Figure 8-4.

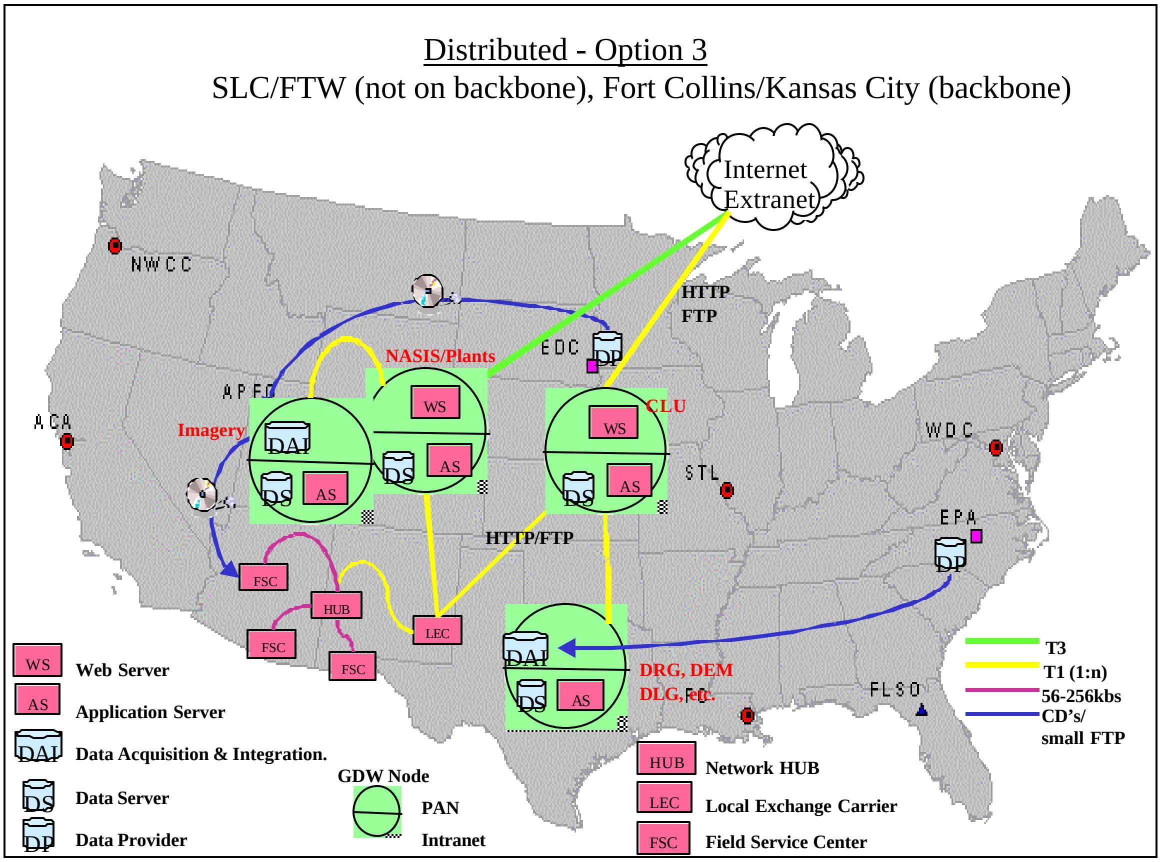


Figure 8-4 – Distributed Option

APFO and NCGC share common functions and continue to disseminate data selectively as restricted by available bandwidth and EA issues. CD-ROM production and delivery continues for those data sets that cannot be disseminated electronically due to size. Data providers continue to supply data sets to the DAIs for integration. In this scenario, all data set ordering through the Gateway is housed at the Fort
Collins web farm. This arrangement fulfills the vision of "one stop shopping" but does not preclude other mechanisms for data retrieval. Data packaging and dissemination functions are still performed at the DAIs. Additionally, APFO and NCGC control the process, hardware, software and applications required to support the Fort Collins web farm or the Kansas City web farm. Table 7-1 outlines the data warehouse functions and where those functions would be managed in a distributed environment.

8.2.3.1. Initial Evaluation

The initial evaluation of this scenario indicates that this distributed option compared with the other distributed options is the most realistic implementation for the short-term. However, the practicality of this implementation needs to be evaluated in terms of performance, therefore the performance of this architecture will be compared to that of the centralized architecture recommended in Section 8.1.2. This option will be piloted through the Gateway/Lighthouse pilot with Fort Collins and NCGC beginning in September 2000.

Pros:

- Data sets are stored and maintained by the data stewards at the point of production, providing one authoritative source.
- Eliminates the cost for USDA to host a single centralized node.
- Promotes extensible framework.

Cons:

- Increased cost of data management components (staffing, hardware, software, and telecommunications).
- If sufficient bandwidth is not available, response time may be impaired due to the number of data stores that may need to be visited per request.
- Each distribution node becomes a single point of failure.

8.2.3.2. Initial Assessment

This scenario has been selected for a more detailed evaluation supported through performance modeling. This scenario was selected due lower implementation costs, as opposed to the scenario described in Section 8.2.2, and the ability to implement from the technological aspect during the FY 2001 time frame. This scenario also supports the long-term USDA vision of distributed data centers hosting non-redundant data sets.

8.2.4. Distributed – OGC/WMT distributed architecture model: internal owned plus external

This is an optimal distributed model, where online geospatial data is shared in an interoperable format between compliant servers located throughout the world. These servers include all levels of government and private/commercial organizations. The model functions according to protocols established and certified by the OpenGIS Consortium. Applications based on this model obtain the data they require from the online sources on an as-needed basis. This eliminates the need to redundantly store data sets locally and enables users to perform functions anytime anywhere without having to download data sets to their local environment. This scenario works very well for those users that often work in disconnect mode in the field. This scenario is directly in line with the Data Team's long-term vision. The OpenGIS specifications are possibly two or more years away, which may be about the time USDA will be prepared to begin the move towards this direction.

8.2.4.1. Initial Evaluation

Pros:

- Data sets are accessible to anyone at anytime and do not need to conform to a common projection system be loaded on a similar platform or utilize a similar GIS system.

Cons:

- The OpenGIS Consortium is responsible for defining the open interfaces and promoting them as a standard for implementation by GIS software vendors. It is up to the vendors themselves to implement the standards and specifications. It may be months to years before this technology is available for implementation.
- This option needs to conform to USDA security measures that may be difficult to enforce.

8.2.4.2. Initial Assessment

This scenario could eventually become a reality for some parts of the USDA business model. USDA should continue to maintain a presence in the Open GIS Consortium to ensure that information on specifications is disseminated within the agency. This architecture is very much in line with the long-term USDA vision should not be removed from consideration for the FY 2003-2005 timeframe. However, at this time implementation is not feasible.

9. Performance Modeling

9.1. Introduction

Performance modeling is a simulation technique that uses mathematical models to predict the performance of a client/server system prior to actually building the system. This technique allows a system designer to construct a model based on a particular architecture and determine how that architecture model will perform based on a set of hardware, software, network components and telecommunications. Once constructed, model parameters can be adjusted and re-simulated to accommodate alternative design considerations. The immediate benefits of modeling include cost and timesaving, risk mitigation, improved decision making and managing scalability.

This modeling effort will be used to evaluate two basic data warehouse alternatives in terms of their ability to handle the anticipated load utilizing the current/planned backbone for FY 2001. If the telecommunications bandwidth is found to be insufficient, a cost estimate to upgrade the telecommunications to an acceptable level will be calculated for that scenario. It is anticipated that the modeling results will support a centralized architecture due to the expected cost to upgrade the telecommunications links between the national centers and the data production centers. Performance results gathered during the modeling will be used in conjunction with the other evaluation criteria to help guide the architecture selection.

Modeling will examine FY 2001 only. However, at some point in the near-term modeling should be performed for FY 2003. FY 2003 modeling should concentrate on system performance if the WAN (Wide Area Network) component is removed from the architecture and users pass directly through UUNet, instead of going through the USDA backbone and then on to UUNet. It is expected that the results of this secondary model may support decentralization. Changes to the recommendations for a FY 2001 architecture need to be considered if modeling supports decentralization in FY 2003.

9.2. Guidelines

Several Data Team meetings and Model Team meetings were held in order to capture the input parameters for the two model designs. The Model Team consists of a subset of Data Team and CCE Team members. In order to model the performance of each scenario, the type and frequency of typical FTP transactions that agencies expect in FY 2001 was constructed. This information is captured in Table 9-1. A separate table that details the type of data access/delivery that should be available for USDA owned datasets is presented in Table 7-2. This table was constructed during a Data Team meeting held at the APFO Field Office in Salt Lake City, UT on July 31, 2000 and August 1, 2000. This table outlines the following data access/delivery methods: read-only online viewing; data streaming; CD-ROM delivery and FTP. Priorities for each access/delivery method are detailed and whether or not the method should be available to the public during the FY 2001 timeframe. The dataset categories modeled by access/delivery type include ortho, soils, CLU, DRGs, NRI, precipitation and “others”.

9.3. Assumptions

Due to the limited amount of time and resources available to build and execute detailed performance models for each of the nine scenarios, only two scenarios were selected for modeling based on rough cost estimates and likelihood of success in FY 2001. The scenarios included a centralized architecture and a distributed architecture. Several assumptions were necessary in order to build the models. These assumptions are presented below in the following categories: general, centralized and distributed.

9.3.1. General Assumptions

- In 2001 all SC data is local. Orders to the Gateway will originate from SCs that have just received GIS and are requesting base data.
- Model will simulate an average workday.
- Data servers and application servers are collocated in order to reduce the amount of time it takes to retrieve data for a requested service.
- Internal and external customers are modeled as one user type. Performance will be based on the bandwidth presently available or available within the FY 2001 timeframe at the SC.
- Files shipped to users are compressed prior to delivery.
- In FY 2001 there will be no data delivery by feature streaming in the prototype.
- Hardware servers selected for model are large enough to handle the anticipated number of requests and processing tasks.
- All application servers are identical in terms of hardware and software composition.
- Performance is based on the current/planned USDA backbone for FY 2001.
- The background load is assumed to be operating at a 75% utilization rate for all telecommunications connections. Complete data could not be obtained for all connections in time for this study. This rate may not accurately portray the actual load on the system. It is expected that the actual rate may be closer to 70-80% utilization on an average basis and 85-90% under peak conditions.
- Hosting the Gateway application in Kansas City instead of Fort Collins will provide better response time for internal users since they will not have to pass through the Fort Collins web farm on the way to the backbone at Kansas City. It is likely that the network topology will shift in FY 2001 as network connections made directly to UUNET.
- The application server located at the Fort Collins web farm or the Kansas City web farm is designated as the central order process server. Software loaded on the application server at Fort Collins or Kansas City contains the Gateway/Lighthouse application, ArcIMS, SDE and navigational services.
- The database server located at Fort Collins or Kansas City is designated as the central ordering database server. This server is assumed present in both scenarios and therefore is not a distinguishing characteristic. This database contains an estimated 2.5 TB of on-line data necessary to support the RDG and other on-line business applications.
- Response times to the end user are not actual response times. The model comparison ends at the WAN. Therefore, SC users and external users are treated as

one client. Although user performance varies according to local connections and the modeling ends at the WAN, user response time is treated as a distinguishing characteristic in the models.

9.3.2. File Size and Transmission Assumptions

Table 9-1 contains estimated file sizes in MB for FY 2001 and FY 2003 ("FY" column) that would constitute typical FTP transactions by APFO and NCGC ("agency" column) and data set ("data" column). These sizes represent the FTP file that is sent in response to a request and can be found in the "product download size MB" column. Estimates were provided by APFO and NCGC. Estimates of the frequency of FTP downloads also provided by APFO and NCGC are found in the "total requests per year" and "total requests per day" columns. Daily frequency numbers are based on 365 days per year. The "up size KB" column contains the estimated size of a typical request sent from a client to a server through a web browser. The "daily web hits" column frequency numbers are based on a single user session on a daily basis and are approximately three times more frequent than FTP downloads. These numbers account for users that are browsing the GDW but not ordering any datasets. Each user is estimated to visit ten web pages in a single session.

The "order request size KB" column in Table 9-1 represents the estimated size of a response sent from the server back to the client in a web browser. These numbers are estimated based on the average size of a response that includes both text and images. This table assumes that FTP and CD production services are maintained at APFO and NCGC. These numbers are most likely conservative, based on current demand and available datasets.

Table 9-1 - Estimated File Sizes and Frequency for FTP Transactions for FY 2001 and FY 2003

agency	data	FY	up size KB	order request size KB	daily web hits	product download size MB	total requests per year	total requests per day
APFO	TIFF[4]	2001	0.5	15	33	200	4,000	11
APFO	SID[5]	2001	0.5	15	33	300	4,000	11
APFO	CLU	2001	0.5	15	82	5	10,000	27
NCGC	SOILS[6]	2001	0.5	15	55	50	6,700	18
NCGC	Other	2001	0.5	15	27	500	3,300	9
APFO	TIFF	2003	0.5	15	49	200	6,000	16
APFO	SID	2003	0.5	15	49	300	6,000	16
APFO	CLU	2003	0.5	15	123	5	15,000	41
NCGC	SOILS	2003	0.5	15	99	50	12,000	33
NCGC	Other	2003	0.5	15	48	500	5,900	16

[4] Uncompressed Ortho Quad
[5] Compressed Ortho for an entire typical sized county
[6] Soil SSURGO data for a typical Soil Survey Area

9.3.3. Model Architecture

The model architecture for the centralized scenario is depicted in Figure 9-1. Two models have been developed for this scenario, one model that places the GDW at the Fort Collins web farm and the other model that places the GDW at the Kansas City web farm. The model will compare the telecommunications infrastructure at each site to determine if one location has an advantage over the other at this time. The front-end data ordering and the back-end data storage and delivery system would all be located at this single node. Therefore, all other parameters, hardware, software and staffing, are constant. Model results for the optimal site will be compared to the distributed model results.

The model architecture for the distributed scenario is depicted in Figure 9-2. This model will look at the difference between hosting the front-end ordering system at either the Kansas City web farm or the Fort Collins web farm and have the data storage and delivery system located at both AFPO and NCGC. Two models will be run in order to determine if Fort Collins or Kansas City would be better equipped to handle the front-end ordering system in a distributed scenario. The back-end data storage and delivery requirements for both APFO and NCGC would be the same regardless of whether Fort Collins or Kansas City was the host of the front-end. Model results for the optimal front-end site will be compared to the optimal results of the centralized scenario to determine the best distribution system scenario, centralized or distributed for FY 2001.

If the performance metrics that come from the centralized model evaluation are considered reasonable, then the cost to create a GDW at APFO or NCGC would be the cost between the current bandwidth at Fort Collins or Kansas City and that of APFO or NCGC.

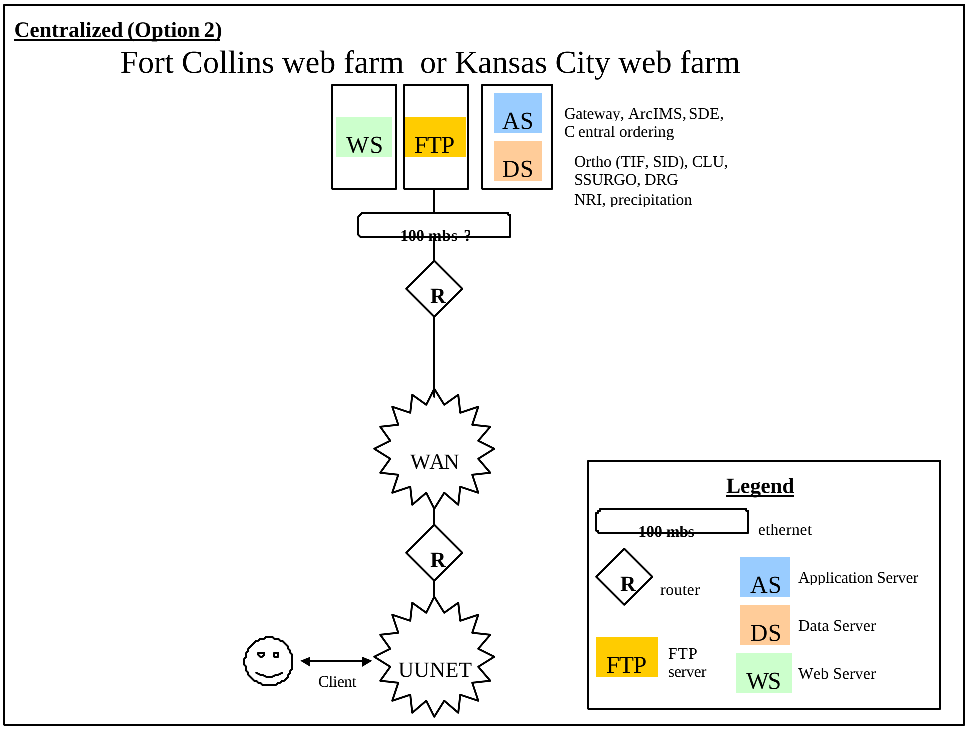


Figure 9-1 – Centralized Scenario Used for Performance Modeling

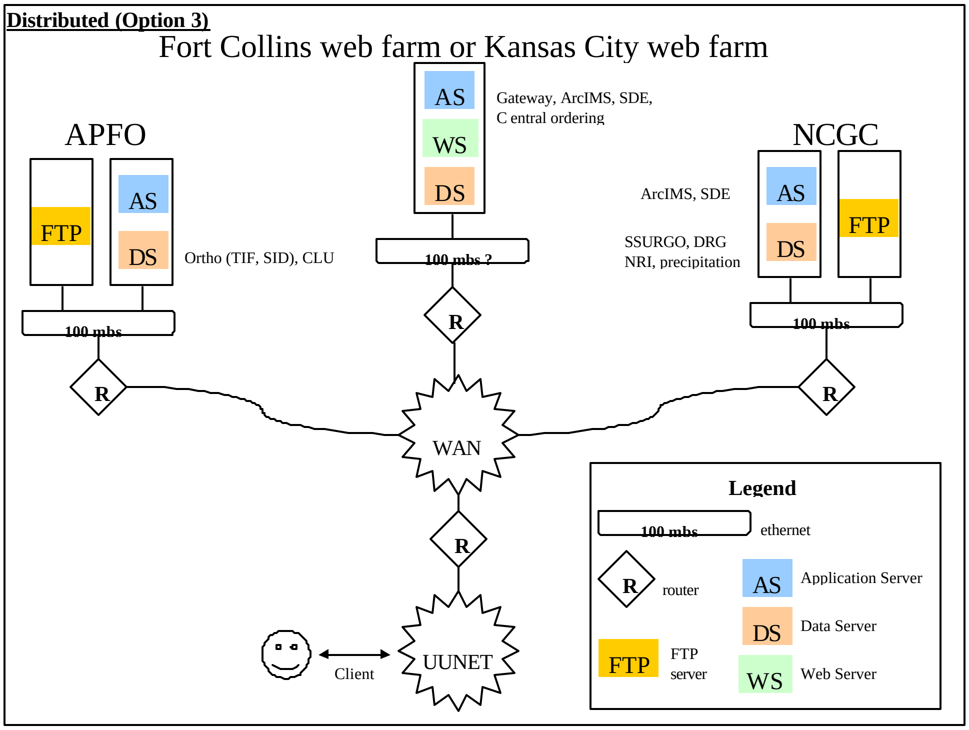


Figure 9-2 – Distributed Scenario Used for Performance Modeling

9.4. Bandwidth Requirements for Performance

Bandwidth requirements were calculated for both the centralized and distributed scenarios in order to compare the current system infrastructure with the expected system demand in FY 2001 and beyond. Minimum and recommended bandwidth requirements were calculated using the estimated file size and transaction frequency metrics listed in Table 9-1. These requirements will help determine the level of telecommunications upgrade that may be necessary if modeling results identify a performance bottleneck in the telecommunications infrastructure between the production centers and the central data warehouse. The cost of an upgrade can be calculated as the cost between recommended bandwidth and the current bandwidth.

The bandwidth currently in place at each facility is represented in the *Current* row. The minimum required bandwidth, calculated from the expected size and frequency of HTTP and FTP transactions, appears in the *Minimum* row. Minimum values are only expected to handle average loads on the system. In order to handle peak loads, recommended bandwidth metrics are listed in the *Recommended (GDW no background load)* row. It is important to note however, recommended bandwidth metrics are based on traffic generated by geospatial data inquiries, orders and deliveries only. These recommendations do not include all other background load that is anticipated on the system, assumed to be at a 75% utilization rate for performance modeling. In order to adequately manage the geospatial traffic and the remainder of USDA business the telecommunication recommendations should be doubled or tripled. This projected estimate is included in the row labeled *Recommended (GDW with background load)*.

Preliminary modeling indicates approximately 10% backbone utilization during average GDW use and approximately 20% during peak operations. All infrastructure connections referred to in this section assume full-duplex lines are utilized. Full-duplex transmission indicates that data can move bi-directionally on a single carrier.

Table 9-2 compares the telecommunications bandwidth differences between the Fort Collins web farm and Kansas City web farm for the centralized architecture scenario. This information will help determine which of the two nodes is best equipped to serve as a central GDW. The current telecommunications listed in the table reflect the bandwidth that extends out from each center to UUNet. The minimum requirements were calculated by adding up the current bandwidth out of APFO (32,000 bits/second), NCGC (1,544,000 bits/second) and the HTTP traffic that would come into either the Fort Collins web farm or Kansas City web farm (2,500 bits/second) and factoring in the expected demand that was previously documented in Table 9-1.

These minimum requirements indicate the level of communication that should be in place during average system access and does not take into consideration peak demand or other background loads placed on the system by USDA business users. The recommended connections only account for load placed on the system by GDW users. Recommended USDA telecommunications are also listed to accommodate both GDW traffic and existing USDA background demands. These requirements are expected for FY 2001 only and should be scaled accordingly for FY 2002 and beyond based on the expected increases in GDW traffic. Traffic is expected to at least double in subsequent years as the system is exposed to more users and USDA increases the number of on-line business applications. A dedicated T-3 line upgrade to the current infrastructure is recommended for a centralized scenario (either located at the Fort Collins web farm or the Kansas City web farm) in order to provide adequate service for the GDW and the expected USDA background, which is likely to exceed the assumed 75% utilization rate (not including GDW traffic) in the near term.

Table 9-2 – Centralized Bandwidth (numbers are in bits/second)

Situation	Fort Collins to UUNet	Kansas City to UUNet
Current	23,160,000 (15 T-1's)	7,720,000 (5 T-1's)
Minimum	2,206,500 (<2 T-1's)	2,206,500 (<2 T-1's)
Recommended (GDW no background load)	5,000,000 (Fractional T3)	5,000,000 (Fractional T3)
Recommended (GDW with background load)	44,736,000 (Dedicated T-3)	44,736,000 (Dedicated T-3)

Table 9-3 documents the current, required and recommended telecommunications bandwidth between the facilities named in each column. Currently, APFO and NCGC are not directly linked to the USDA backbone. APFO traffic is routed through Fort

Collins and then sent on to Kansas City. NCGC traffic is routed through WDC before going on to Kansas City. Both Fort Collins and Washington DC (WDC) have 2 T-1's that connect to the Kansas City web farm. Additionally, NCGC currently shares its T-1 line (1,544,000 bits/second) to WDC with 23 Field Service Centers. NCGC is planning to increase the number of T-1 connections to a total of three in FY 2001. One of these T-1's will be dedicated to geospatial data delivery.

In order to handle the expected GDW traffic, two T-1 lines are recommended for each APFO and NCGC, however the recommendation for both GDW and USDA traffic is two-thirds of a T-3 line. As demonstrated by the table, the existing Kansas City web farm and Fort Collins web farm configurations meet the recommended GDW specifications but will need a partial upgrade to handle both GDW and USDA traffic.

Another option for FY 2001, is to upgrade the telecommunications infrastructure between APFO and NCGC directly to UUNet as opposed to upgrading the existing lines currently in place and depicted in the table. If this option were selected the size of the lines between APFO and NCGC to UUNet are estimated to be the same as those calculated for APFO to the Fort Collins web farm and NCGC to WDC or the Kansas City web farm or Fort Collins web farm. This option may be more cost effective in the long run as the EAI moves towards replacing the USDA backbone with an UUNet backbone.

Table 9-3 – Distributed Bandwidth (numbers are in bits/second)

Situation	APFO to Fort Collins (or UUNet)	NCGC to WDC (or UUNet)	Kansas City to UUNet	Fort Collins to UUNet
Current	32,000 (N/A)	1,544,000 (T-1)	7,720,000 (5 T-1's)	23,160,000 (15 T-1's)
Minimum	1,004,000 (< 1 T-1)	1,200,000 (< 1 T-1)	2,500 (N/A)	2,500 (N/A)
Recommended (GDW no background load)	3,088,000 (2 T-1 lines to UUNet)	3,088,000 (2 T-1 lines to UUNet)	772,000 (½ T-1 line)	772,000 (½ T-1 line)
Recommended (GDW with background load)	29,525,760 (2/3 T-3 line to UUNet)	29,525,760 (2/3 T-3 line to UUNet)	29,525,760 (2/3 T-3 line)	29,525,760 (2/3 T-3 line)

10. Evaluation and Recommendation

To present a set of recommendations for short and long-term geospatial data warehouse implementation each of the candidate scenarios selected for further analysis were weighted according to the evaluation criteria presented in Section 7.4. This section presents the results of the performance modeling and address each of the evaluation criteria in more detail for both the centralized and distributed scenarios.

10.1. Performance Modeling Results

A centralized model was developed to determine whether or not a centralized web farm, located at either Fort Collins or Kansas City would support the level of traffic expected at the GDW in FY 2001. The simulation examined the bandwidth utilization for a single centralized web farm having 15 T-1 (23,160,000 bps) lines connected to UUNet. A second simulation increased the telecommunications slightly to a Fractional T-3 line (29,525.760 bps). The results indicate that for either telecommunications scenario, the utilization rate is 84.2%, assuming 75% of that utilization was background load. This indicates that the GDW is adding a 9.2% load to the existing assumed 75% utilization. The response time for the centralized approach is 1.84 seconds which is higher than that of the distributed model but may still be within reason for FTP based transactions.

Two distributed models were developed in order to compare the bandwidth performance between the Fort Collins web farm and the Kansas City web farm based on the existing telecommunications infrastructure. The telecommunications connections between the central web farm and the distributed centers are based on existing connections. The performance results indicate that for each of the possible web farm locations the 32,000 bps link between APFO and is saturated at 98.5% utilization in the Fort Collins web farm scenario and at 100% utilization for the Kansas City web farm. These utilization values confirm the need to upgrade the telecommunications infrastructure at APFO. The saturation at APFO occurred during the first ten minutes of running the model, thus suspending the simulation. Since the model could not complete, utilization values other than APFO and responses times are not going to provide an accurate picture of the system performance.

A third distributed model was developed in order to test bandwidth performance of a distributed system in FY 2003 where all components have direct connections to UUNet. In this simulation APFO, NCGC, Fort Collins web farm and Kansas City web farms were assumed to have a Fractional T-3 connection (29,525,760 bps) to UUNet. This simulation showed a reduction in the utilization of the telecommunications infrastructure between APFO and UUNet, from saturation (existing telecommunications) to 87.5% (Fractional T-3) and an 82.7% (Fractional T-3) utilization between NCGC and UUNet. The overall response time for this simulation was 0.154 seconds. The telecommunications between UUNet and the Fort Collins web farm and UUNet and the Kansas City web farm are equal, thus eliminating web farm location as an indicator of performance.

10.2. Scenario Cost Analysis

Table 10-1 is a summary cost analysis performed on the centralized and distributed architecture scenarios. During the development of the scenarios and the performance modeling, four scenario architectures were presented. These were:

1. Centralized Kansas City web farm
2. Centralized Fort Collins web farm
3. Distributed Kansas City web farm, APFO, NCGC

4. Distributed Fort Collins web farm, APFO, NCGC

After review and analysis, it was decided to combine the analysis for the Kansas City web farm and Fort Collins web farm into a generic Centralized Web Farm. This effectively reduced the cost analysis to two scenarios. These are:

1. Centralized - Web Farm
2. Distributed - Web Farm, APFO, NCGC

Table 10-1 FY 01 Summary Costs for Centralized Geospatial Data Warehouse (New)

Centralized (New)		
Item	**Centralized Web Farm (New Resources Only)**	**Sub Total New**
Hardware	$489,420	
Software	$0	
Network	$388,677	
Staff	$782,781	
Total	$1,660,878	**$1,660,878**

Table 10-2 FY 01 Summary Costs for Distributed Geospatial Data Warehouse (New)

Distributed (New)				
Item	**Distributed Web Farm (For New and Upgrade)**	**Distributed APFO (New)**	**Distributed NCGC (New)**	**Total Distributed (New)**
Hardware	$607,951	$753,011	$828,011	
Software	$0	$0	$0	
Network	$388,677	$384,765	$377,346	
Staff	$971,333	$1,967,206	$1,967,206	
Total	$1,967,961	$3,104,982	$3,172,563	**$8,245,506**

Table 10-3 FY 01 Summary Costs for Distributed Geospatial Data Warehouse (Existing)

Distributed (Upgrade from Existing)				
Item	**Distributed Web Farm (For New and Upgrade)**	**Distributed APFO (Upgrade)**	**Distributed NCGC (Upgrade)**	**Total Distributed (Upgrade from Existing)**
Hardware	$607,951	$292,458	$341,308	
Software	$0	$0	$0	
Network	$388,677	$384,765	$384,765	
Staff	$971,333	$258,014	$39,811	
Total	$1,967,961	$935,237	$765,884	**$3,669,082**

10.2.1. Scenario Cost Assumptions

New Resources vs. Upgrade from Existing Resources - The cost development process revealed that many resources currently exist that could be applied directly to or upgraded to fit the proposed architecture. These resources, both human and capital, were mainly assets of the two exiting Data Acquisition, Integration and Delivery centers (APFO and NCGC). No existing assets were identified from the Web Farm in either the Centralized or Distributed scenario. Table 10-1 itemized costs for both new and upgraded resources.

Hardware - Hardware costs included the following server and peripheral equipment:

- Web Server
- Application Server
- Data Server
- FTP Server
- Tape Backup
- Online Storage
- CD-ROM Production

Make and model of the selected hardware was based on the preferred or existing vendor for each facility. Size and capacity was based on the performance modeling results, comparison with similar systems and experience.

Software - At the time of this draft, no software costs were provided. Both ESRI and Microsoft are negotiating enterprise license cost for their software. A later draft may include GSA list price if this is not resolved.

Network - The network sizing requirements were built up from the performance modeling done by American Management Systems. A network cost matrix was developed by the Fort Collins Information Technology Center. Router costs were not available during the time of this draft.

Staff - Staff position descriptions were developed as a product of this study. FTE equivalents were assigned for each position/location in each scenario. Understaffed positions were priced as required staff. Overstaffed positions were priced at zero cost. Under either scenario, data management would be the responsibilities of the data stewards.

10.2.2. Scenario Cost Assessment

This costs assessment is based on preliminary analysis and the assumptions in the section above. The assessment shows the following:

1. Upgrading from existing resources has a significant impact on leveling the cost between centralized and decentralized options.
2. The telecommunications infrastructure for each site is so inadequate that the cost differential between upgrading the existing connections or installing new connections becomes insignificant.
3. Table 7-2 lists the total storage requirement for Service Center Imagery as approximately 24 TB. It is assumed that would not require that total storage amount for the first year, but ramp up to total storage capacity over several years. The

following table shows the cost of ownership over 5 years. Estimates of the cumulative storage requirements are based on projected growth of the database. The annual investment is based on a 50% annual price reduction starting at $100,000/TB in FY 01.

Table 10-4 Sample cost of ownership matrix for on-line storage

Year	FY 01	FY 02	FY 03	FY 04	FY 05
Cost/TB	100000	50000	25000	12500	6250
Incremental Increase (TB)	2	4	4	6	8
Annual Cost	$200,000	$200,000	$100,000	$75,000	$50,000

10.3. Near-term Recommendations (FY 2001)

10.3.1. Automate FTP and CD-ROM ordering and delivery

Continue to deliver data to the Service Centers, partners and customers using FTP when file sizes are within a reasonable download time. Continue CD-ROM data delivery for the initial base set of data and any additional data sets that cannot be reasonably transferred given the current state of telecommunications.

10.3.2. Invest in on-line data services

Pilot and expand data streaming delivery as business, technology and telecommunications permits. Data streaming will reduce the amount of unnecessary redundant storage as well as provide more current data to the users. Under the distributed and centralized options data warehouses and data marts would be built support of on-line data services

10.3.3. Invest in the telecommunications infrastructure between APFO and NCGC

Invest in the telecommunications infrastructure between APFO, NCGC and USDA Web Farms. This investment will support the management and transfer of datasets between the production centers and the GDW in a centralized architecture or between the production centers and their FTP clients in a distributed architecture. Invest in UUNet or other carrier connections between DAIs, the Internet and the USDA backbone.

10.3.4. Exploit Existing IT Resources

Where possible, use or upgrade existing hardware and software in each of the proposed GDW locations. Capitalize on existing staff resources and contract supplemental staff to support GDW operations.

10.3.5. Prioritize Data Access and Delivery to USDA

Make data access and delivery to USDA users the primary priority. Public and partner access should be secondary. Exceptions should be made when public access is mandated or cost sharing could be accomplished. Implement public access solutions as time and budget permit. However, all scenarios must consider public access in their implementation plans in order to provide for growth in this area of delivery.

10.3.6. Continue Technology Insertion

Continue to actively infuse new technology into the architecture framework as it becomes available. Maintain active participation in the OpenGIS Consortium work with the Web Mapping Testbed.

10.4. Long-term Recommendations (FY 2002 and beyond)

10.4.1. Decrease Dependency on CD-ROM and increase electronic data delivery

As bandwidth increases are implemented according to the plans laid out by EA and CCE, decrease the number of CD-ROM deliveries and increase FTP transactions.

10.4.2. Foster Data Partnerships

Continue to pursue data storage and data delivery partnership agreements with other land managing agencies.

10.4.3. Increase Public Access Capability When Mandated

Phase in alternative access solutions for the public including online viewing and downloading as permitted by security policies.

Appendix A – Bibliography

[A1] USDA Service Center Geographic Information Systems (GIS) Strategy, August 1998

[A2] Geospatial Data Acquisition, Integration and Delivery National Implementation Strategy Plan, September 1999

[A3] Geospatial Data Requirements, April 2000

[A4] Standard for Geospatial Data, January 2000

[A5] Standard for Geospatial Data Set Metadata, August 1999

[A6] Standard for Geospatial Dataset File Naming, August 2000

[A7] Standard for Service Center Tabular Metadata, September 1999

[A8] Service Center Data Administration Concept of Operations, August 1998

[A9] Executive Order 12905, Published in the April 13, 1994, edition of the Federal Register, Volume 59, Number 71, pp. 17671-17674

[A10] The OpenGIS Abstract Specification Topic 13: Catalog Services Version 4, 1999